

Expandable Factor Analysis

Sanvesh Srivastava ^{*1,2}, Barbara E. Engelhardt ^{†1} and David B. Dunson ^{‡1}

¹*Department of Statistical Science, Duke University, Durham, NC*

²*Statistical and Applied Mathematical Sciences Institute, Durham, NC*

December 10, 2014

Abstract

Bayesian sparse factor models have proven useful for characterizing dependence, but scaling computation to high dimensions is problematic. We propose *expandable factor analysis* for scalable estimation. The method relies on a novel *multiscale generalized double Pareto* shrinkage prior that allows efficient estimation of low-rank and sparse loadings matrices through weighted ℓ_1 -regularized regression. Integrated nested Laplace approximations are used for model averaging to accommodate uncertainty in the number of factors and tuning parameters. We provide theoretical support and develop efficient computational algorithms. The methods are applied to simulated data and genomic studies.

Key words: Bayesian model averaging; Expectation-Maximization-type algorithm; factor analysis; generalized double Pareto; high-dimensional loadings matrix; integrated nested Laplace approximation; non-concave variable selection; sparsity.

1 Introduction

Latent factor models provide a useful approach for estimating covariance matrices, but face challenges in scaling to high-dimensional data. Letting Ω denote an unknown $P \times P$ covariance matrix, the usual factor model assumes $\Omega = \Lambda\Lambda^T + \Sigma$, with Λ a $P \times K$ *loadings* matrix and Σ a $P \times P$ diagonal residual variance matrix. To allow estimation of Ω in cases where the sample size N is small relative to the dimension P , one can impose low-rank and sparsity assumptions on Λ . A low-rank Λ has K much smaller than P , and a sparse Λ has few non-zero loadings. Bayesian approaches to sparse factor analysis are appealing in allowing uncertainty in both K and the locations of the zero elements (Carvalho et al., 2008; Knowles and Ghahramani, 2011). However, in large P settings, it is intractable to draw samples from the posterior distribution for K and the locations of zeros. Applying a continuous shrinkage prior that favors many values near zero,

^{*}ss602@stat.duke.edu

[†]barbara.engelhardt@duke.edu

[‡]dunson@duke.edu

while adaptively selecting K , provides a partial solution to the computational bottleneck (Bhattacharya and Dunson, 2011; Pati et al., 2014).

Penalized optimization methods provide an efficient alternative to obtain point estimates of Λ and Σ . If K is assumed to be known, then many such methods exist (Rubin and Thayer, 1982; Bai and Li, 2012; Hirose and Yamamoto, 2014). For example, sparse principal components analysis estimates a sparse loadings matrix assuming $\Sigma = \sigma^2 I_P$, with I_P the $P \times P$ identity matrix (Jolliffe et al., 2003; Zou et al., 2006; Shen and Huang, 2008; Witten et al., 2009). The assumptions of spherical residual covariance and known K are restrictive in practice. There are several approaches available to estimate K . In econometrics, it is popular to rely on test statistics based on the eigenvalues of the empirical covariance (Lam and Yao, 2012; Ahn and Horenstein, 2013). Alternatives rely on fitting the model for a variety of K values, and choosing the best value based on a model selection criterion. In contrast to the rich literature studying such an approach in penalized regression, the behavior of such approaches is largely unknown in the factor analysis setting (Chen and Chen, 2008; Hirose and Yamamoto, 2014).

We are motivated to limit sensitivity to choice of tuning parameters, while also providing a characterization of uncertainty. With these goals, we propose an approximate Bayesian approach that relies on a novel multiscale generalized double Pareto prior. This prior is inspired by the generalized double Pareto prior for variable selection (Armagan et al., 2013), and by the multiplicative gamma process prior for loadings matrices (Bhattacharya and Dunson, 2011). The latter approach focuses on estimation of Ω , and cannot be used to estimate Λ due to the lack of identifiability constraints.

Multiscale generalized double Pareto priors are used to address uncertainty in K , efficient estimation of low-rank and sparse Λ , and model uncertainty. A local linear approximation of the multiscale generalized double Pareto penalty leads to efficient maximum a posteriori estimation of Λ using weighted ℓ_1 -regularized regression (Zou and Li, 2008). The carefully structured prior ensures that the estimated Λ has rank $K = \mathcal{O}(\log P)$ and leads to analytically tractable posterior model weights using integrated nested Laplace approximation. We handle uncertainty due to prior specification by averaging Λ estimates across a grid of hyperparameters using posterior model weights. Expandable factor analysis combines the strengths of Bayesian and penalized likelihood methods for factor analysis, while having guarantees for consistent parameter estimation under standard regularity conditions of variable selection.

2 Expandable Factor Analysis

2.1 Factor analysis as regression

Consider the usual factor analysis model for Gaussian response

$$y_n = \Lambda z_n + e_n, \quad z_n \sim \mathcal{N}(0, I_K), \quad e_n \sim \mathcal{N}(0, \Sigma), \quad n = 1, \dots, N. \quad (1)$$

Denote the data matrix centered at sample mean as $Y = [y_1, \dots, y_N]^T$, the factor matrix as $Z = [z_1, \dots, z_N]^T$, and the residual matrix as $E = [e_1, \dots, e_N]^T$. Factor analysis is equivalent to P regressions

in the factor space, so that (1) reduces to

$$y_p = Z\lambda_p + e_p, \quad e_p \sim \mathcal{N}(0, \sigma_{pp}^2 I_N), \quad p = 1, \dots, P. \quad (2)$$

The crucial difference between factor analysis and regression is that the design matrix Z is unknown.

Current approaches for penalized factor analysis can be understood as regularized estimation of λ_p s in (2), with different choices of penalty. Hirose and Yamamoto (2014) propose one such approach by regularizing λ_p s using MC+ penalty (Zhang, 2010a). Their loss is based on the log likelihood obtained from standard Expectation-Maximization arguments for factor analysis. Reinterpreting their computational algorithm using (2) shows that if K is fixed a priori, then they iteratively estimate Λ through P MC+-regularized linear regressions.

Expandable factor analysis is motivated by high-dimensional genomic applications in which K is unknown. A Bayesian approach models this uncertainty by using a prior for Λ with support in $\mathbb{R}^{P \times \infty}$. The prior also regularizes Λ by putting high probabilities on neighborhoods of Λ with $K \ll P$. Such a Bayesian approach to factor analysis is described next.

2.2 Desirable properties of a prior

Taking a Bayesian approach, one can specify a prior on Λ , with K ranging from 1 to ∞ and the loadings increasing shrunk towards zero as column increases. Such priors are conceptually related to stick-breaking priors used in mixture models, which do not require specification of a finite number of mixture components (Sethuraman, 1994). Many variations of spike-and-slab and shrinkage priors exist for factor analysis, but current approaches are computationally intractable in high-dimensions and do not provide theoretical guarantees for consistent estimation of Λ (Carvalho et al., 2008; Bhattacharya and Dunson, 2011; Knowles and Ghahramani, 2011). We propose to solve these problems using an approximate Bayesian approach.

We identify four properties of a good prior or penalty for factor analysis. Three of the four properties follow from non-concave variable selection (Fan and Li, 2001). They are reinterpreted as follows in the context of factor analysis:

- (a) *Near unbiasedness*: If an element λ_{ij} of the true and unknown loadings matrix is large, then its estimate $\hat{\lambda}_{ij}$ should be nearly unbiased.
- (b) *Sparsity*: The estimate $\hat{\lambda}_{ij}$ of the loading λ_{ij} is obtained using a thresholding rule, such as soft-thresholding. This rule ensures that small $\hat{\lambda}_{ij}$ s are automatically set to zero.
- (c) *Continuity*: The estimator of Λ is continuous in data Y to limit instability.
- (d) *Ordered factors*: The columns of Λ are ordered such that the loadings are increasingly shrunk towards zero as K increases from 1 to ∞ and the total number of non-zero loadings decreases across columns of Λ .

Properties (b) and (d) together ensure existence of a finite column after which all elements of the estimated Λ are identically zero.

Whether the priors satisfy properties (a) – (c) depends in part on what type of estimator $\hat{\lambda}_{ij}$ is used. From a Bayesian perspective, $\hat{\lambda}_{ij}$ should correspond to the value minimizing the Bayes risk, which is the expectation of a loss function averaged over the posterior for λ_{ij} . In general, the posterior mean, which corresponds to the Bayes estimator under squared error loss, will not be sparse even if the prior allows zero values with positive probability. Potentially, we could define a more elaborate loss function that induces sparse estimates even under a continuous shrinkage prior, which does not allow exact zeros. However, from a practical perspective, it may be intractable to compute the Bayes estimator under such loss functions; even the posterior mean is typically intractable as dimensionality increases. For this reason, we focus on $\hat{\lambda}_{ij}$ corresponding to the posterior mode; for this choice, none of the existing priors satisfy properties (a)–(d) simultaneously, while the prior proposed in the next subsection does.

2.3 Multiscale generalized double Pareto prior

Bhattacharya and Dunson (2011) show that the set of all loadings matrices, $\Lambda \in \mathbb{R}^{P \times \infty}$, that lead to well-defined covariance matrices corresponds to

$$\mathcal{C}_{\text{load}} = \left\{ \Lambda = (\lambda_{pk}, p = 1, \dots, P, k = 1, \dots, \infty) : \max_{1 \leq p \leq P} \sum_{k=1}^{\infty} \lambda_{pk}^2 < \infty \right\}. \quad (3)$$

We propose a multiscale generalized double Pareto prior for Λ having support on $\mathcal{C}_{\text{load}}$. This prior is carefully chosen to concentrate near low-rank and sparse matrices, placing high probability around matrices having rank $\mathcal{O}(\log P)$.

The multiscale generalized double Pareto prior specifies independent generalized double Pareto priors on elements of λ_p such that property (d) is satisfied. Specifically,

$$p_{\text{mGDP}}(\lambda_p) = \prod_{k=1}^{\infty} p_{\text{GDP}}(\lambda_{pk} | \alpha_k, \eta_k) = \prod_{k=1}^{\infty} \frac{\alpha_k}{2\eta_k} \left(1 + \frac{|\lambda_{pk}|}{\eta_k} \right)^{-(\alpha_k+1)} \quad (4)$$

for $p = 1, \dots, P$, and $p_{\text{mGDP}}(\Lambda) = \prod_{p=1}^P p_{\text{mGDP}}(\lambda_p) \equiv \text{mGDP}(\alpha_{1:\infty}, \eta_{1:\infty})$. The $\text{GDP}(\alpha_k, \eta_k)$ prior on λ_{pk} ensures that properties (a) – (c) are satisfied. Property (d) is satisfied by choosing parameter sequences $\alpha_{1:\infty}$ and $\eta_{1:\infty}$ such that two conditions hold: the prior measure \mathbb{P}_{load} induced by $\text{mGDP}(\alpha_{1:\infty}, \eta_{1:\infty})$ on $\mathcal{C}_{\text{load}}$ is a valid probability measure and \mathbb{P}_{load} has $\mathcal{C}_{\text{load}}$ as its support. The following lemma identifies the conditions on $\alpha_{1:\infty}$ and $\eta_{1:\infty}$ such that $\mathbb{P}_{\text{load}}\{\Lambda : \Lambda \in \mathcal{C}_{\text{load}}\} \equiv \mathbb{P}_{\text{load}}\{\mathcal{C}_{\text{load}}\} = 1$.

Lemma 2.1 *If $\alpha_k > 2$ and $\frac{\eta_k}{\alpha_k} = \mathcal{O}\left(\frac{1}{k^m}\right)$ $k = 1, \dots, \infty$ and $m > 0.5$, then $\mathbb{P}_{\text{load}}\{\mathcal{C}_{\text{load}}\} = 1$.*

Remark 2.1 *The sequence $\{\eta_k/\alpha_k\}_{k=1}^{\infty}$ is decreasing and $\alpha_k > \eta_k > 0$ for $k = 1, \dots, \infty$. We choose a particular form of $\alpha_{1:\infty}$ and $\eta_{1:\infty}$ based on computational convenience. The parameters $\alpha_{1:\infty}$ and $\eta_{1:\infty}$ also depend on N to guarantee consistency of estimated Λ ; see Theorem 4.2.*

Lemma 2.1 is too general to be practically useful for two reasons. First, $K = \infty$ in the multiscale generalized double Pareto prior does not lead to feasible computations. We resolve this limitation, following

Bhattacharya and Dunson (2011), by mapping any $\Lambda \in \mathcal{C}_{\text{load}}$ to $\Lambda^{K_0} \in \mathcal{C}_{\text{load}}$ that retains only the first K_0 columns of Λ . We choose K_0 such that $\Omega^{K_0} = \Lambda^{K_0} \Lambda^{K_0^\top} + \Sigma$ is arbitrarily close to Ω , where the distance between Ω^{K_0} and Ω is measured as ℓ_∞ norm of their element-wise difference. The second limitation is that learning tuning parameters $\alpha_{1:K_0}$ and $\eta_{1:K_0}$ is practically infeasible. We address this limitation by defining $\alpha_{1:K_0}$ and $\eta_{1:K_0}$ as functions of tuning parameters δ and ρ , respectively, such that property (d) is satisfied. These functions and the approximate number of factors K_0 are defined by the following lemma such that Lemma 2.1 holds.

Lemma 2.2 $\mathbb{P}_{\text{load}}\{\mathcal{C}_{\text{load}}\} = 1$ holds when any one of the following three sufficient conditions are satisfied. For $k = 1, \dots, \infty$,

(I) $\alpha_k = \delta^k$ and $\eta_k = \rho$ where $\delta > 2$ and $\rho > 0$;

(II) $\alpha_k = \delta$ and $\eta_k = \rho^k$ where $\delta > 2$ and $1 > \rho > 0$;

(III) $\alpha_k = \delta^k$ and $\eta_k = \rho^k$ where $\delta > 2$ and $\delta > \rho > 0$.

Furthermore, given $\epsilon > 0$ and depending on which one of the three sufficient conditions (I), (II), and (III) is satisfied, for every Ω , there exists $K_0(P, \rho, \delta, \epsilon)$ such that for all $K \geq K_0$ and $\Omega^K = \Lambda^K \Lambda^{K^\top} + \Sigma$, $\mathbb{P}\{\Omega^K \mid d_\infty(\Omega, \Omega^K) < \epsilon\} > 1 - \epsilon$, where $d_\infty(A, B) = \max_{1 \leq i, j \leq P} |a_{ij} - b_{ij}|$. If (I) holds, then $K_0 = \mathcal{O}(\log^{-1} \delta \log \frac{P}{\epsilon^2})$; if (II) holds, then $K_0 = \mathcal{O}(\log^{-1} \frac{1}{\rho} \log \frac{P}{\epsilon^2})$; and if (III) holds, then $K_0 = \mathcal{O}(\log^{-1} \frac{\delta}{\rho} \log \frac{P}{\epsilon^2})$.

We use only those $\text{mGDP}(\alpha_{1:\infty}(\delta), \eta_{1:\infty}(\rho))$ priors for factor analysis that satisfy one of the three sufficient conditions (I) – (III).

3 Estimation of loadings and residual variance matrices

3.1 Parameter estimation in expandable factor analysis

We use the multiscale generalized double Pareto prior for Λ , Jeffrey's prior for σ_{pp}^2 for $p = 1, \dots, P$, and fix K at a conservative upper bound for K_0 in Lemma 2.2. Estimation of Λ and Σ reduces to P separate regularized regressions. If $\Lambda^{(t)}$ and $\Sigma^{(t)}$ are the estimates at the t th iteration, then the next parameter updates depend on the data through $\Psi^{(t)} = E(Z^\top Z \mid Y, \Lambda^{(t)}, \Sigma^{(t)})/N$ and $\hat{\Lambda}^{(t)} = E(\sum_{n=1}^N y_n z_n^\top \mid Y, \Lambda^{(t)}, \Sigma^{(t)})/N$; see Appendix B.1. Defining pseudo response $w_p^{(t)} = \Psi^{(t)-1/2} \hat{\Lambda}_p^{(t)}$ and pseudo design matrix $X^{(t)} = \Psi^{(t)1/2}$ leads to estimation of Λ and Σ in the $(t+1)$ th iteration as

$$\begin{aligned} \underset{\lambda_p, \sigma_{pp}^2}{\text{argmin}} \quad & \frac{N+2}{2} \log \sigma_{pp}^2 + \frac{N}{2} \frac{\|w_p^{(t)} - X^{(t)} \lambda_p\|^2 - w_p^{(t)\top} w_p^{(t)} + (Y^\top Y/N)_{pp}}{\sigma_{pp}^2} \\ & + \sum_{k=1}^K (\alpha_k + 1) \log \left(1 + \frac{|\lambda_{pk}|}{\eta_k} \right) \end{aligned} \quad (5)$$

for $p = 1, \dots, P$; see Appendix B.1 for details.

The objective (5) is non-convex in $(\lambda_p, \sigma_{pp}^2)$, but it is convex in $1/\sigma_{pp}^2$ given λ_p . We fix λ_p at $\lambda_p^{(t)}$ in (5) and maximize (5) to obtain the $(t+1)$ th update of σ_{pp}^2

$$\sigma_{pp}^{2(t+1)} = \frac{N}{N+2} \left\{ (Y^T Y / N)_{pp} + \lambda_p^{(t)\top} \Psi^{(t)} \lambda_p^{(t)} - 2 \hat{\lambda}_p^{(t)\top} \lambda_p^{(t)} \right\}. \quad (6)$$

The next subsection modifies (5) to estimate $\lambda_p^{(t+1)}$ given $\sigma_{pp}^{2(t)}$ by solving a convex program.

3.2 Estimation of loadings matrix through a convex objective function

Fixing σ_{pp}^2 at $\sigma_{pp}^{2(t)}$ and replacing the multiscale generalized double Pareto penalty by its local linear approximation in (5) reduces the estimation of λ_p at the $(t+1)$ th iteration to

$$\lambda_p^{\text{lla}(t+1)} = \underset{\lambda_p}{\operatorname{argmin}} \frac{N}{2} \frac{\|w_p^{(t)} - X^{(t)} \lambda_p\|^2}{\sigma_{pp}^{2(t)}} + \sum_{k=1}^K \frac{\alpha_k + 1}{\eta_k + |\lambda_{pk}^{(t)}|} |\lambda_{pk}|. \quad (7)$$

The objective (7) corresponds to a regularized regression with response $w_p^{(t)}$, design matrix $X^{(t)}$, regression coefficients λ_p , and weighted ℓ_1 penalty on λ_p . Its optimum is obtained using block coordinate descent. Representing k th column of Ψ and p th row of Λ without their k th elements as $\Psi_{(-k),k}$ and $\lambda_{p,(-k)}^T$ yields the following single step of coordinate descent update for λ_p^{lla}

$$\lambda_{pk}^{\text{lla}(t+1)} = \frac{\operatorname{sign}(\tilde{\lambda}_{pk}^{(t)})}{\Psi_{kk}^{(t)}} \left(|\tilde{\lambda}_{pk}^{(t)}| - c_{pk}^{(t)} \right)_+, \quad c_{pk}^{(t)} = \frac{(\alpha_k + 1) \sigma_{pp}^{2(t)}}{N(\eta_k + |\lambda_{pk}^{(t)}|)}, \quad k = 1, \dots, K, \quad (8)$$

$\tilde{\lambda}_{pk}^{(t)} = \hat{\lambda}_{pk} - \lambda_{p,(-k)}^{\text{lla}(t)\top} \Psi_{(-k),k}^{(t)}$, and $(\cdot)_+$ is the soft-thresholding operator. We also exploit the structure of (8) and estimate Λ by successively updating each of its columns $\lambda_k^{\text{lla}(t)}$ starting from $k = 1$ to K ; see Appendix B.2 for details.

The Λ estimate obtained using (8) satisfies properties (a) – (c). The adaptive threshold $c_{pk}^{(t)}$ in (8) ensures that property (a) is satisfied. The soft-thresholding rule for estimation of λ_{pk} ensures that property (b) is satisfied. Because both (6) and (7) have continuous first derivatives in the parameter space excluding zero, property (c) is also satisfied (Zou and Li, 2008). Additionally, the estimated Λ satisfies property (d) due to the structure of the multiscale generalized double Pareto penalty.

The efficiency of block coordinate descent is improved further using one-step estimation. If we replace $|\lambda_{pk}^{(t)}|$ in (7) by any \sqrt{N} -consistent estimate of λ_{pk} for $k = 1, \dots, K$, it leads to a consistent estimator of Λ in a single step of coordinate descent; see Van der Vaart (2000) for theoretical details of one-step estimation and Zou and Li (2008) for its application in penalized regression. One-step estimation also improves the theoretical convergence properties. While any \sqrt{N} -consistent estimates of Λ and Σ could be used, we use the maximum likelihood estimates as \sqrt{N} -consistent estimates based on the theoretical results of Bai and Li (2012) in high-dimensional factor analysis.

The estimate of Λ depends on the choice of tuning parameters ρ and δ . Current methods construct a solution surface by varying the tuning parameters on a grid and choosing the values yielding the best model

selection criterion. Results based on such an approach are often unstable, with sensitivity to choice of grid, minor perturbations of the data and criteria used, without good justification for one criteria over others. We propose a Bayesian model averaging approach to reduce these problems.

3.3 Bayesian model averaging for expandable factor analysis

We use Bayesian model averaging to accommodate uncertainty in Λ due to hyper-parameter choice. Every coordinate (ρ_g, δ_g) of a ρ - δ grid corresponds to a factor analysis model \mathcal{M}_g with $\text{mGDP}(\alpha_{1:K}^g, \eta_{1:K}^g)$ prior and parameter estimates Λ^g and Σ^g . Assume that there are G grid indices $g = 1, \dots, G$. If $\pi_g = \mathbb{P}(\mathcal{M}_g | Y)$ and $p(\mathcal{M}_g)$ are, respectively, the posterior and prior probability of \mathcal{M}_g , then the model averaged loadings matrix estimate is

$$\bar{\Lambda}^{\text{lla}} = \sum_{g=1}^G \pi_g \Lambda^g, \quad \pi_g = \frac{p(Y | \mathcal{M}_g) p(\mathcal{M}_g)}{\sum_{g=1}^G p(Y | \mathcal{M}_g) p(\mathcal{M}_g)}; \quad (9)$$

the marginal likelihood

$$p(Y | \mathcal{M}_g) = \int p(Y | \Lambda, \Sigma, \mathcal{M}_g) p(\Lambda) p(\Sigma) d\Lambda d\Sigma \quad (10)$$

is analytically intractable. The estimate $\bar{\Lambda}^{\text{lla}}$ (9) averages loadings matrix estimates $\Lambda_{\rho, \delta}$ s across the grid of hyper-parameter values, with the weights corresponding to posterior probabilities under a uniform prior over the grid.

We approximate (10) to obtain an analytic form for $p(Y | \mathcal{M}_g)$, and in turn π_g and $\bar{\Lambda}^{\text{lla}}$. The values of (ρ_g, δ_g) only affect Λ^g , so we first replace $p(Y | \Lambda, \Sigma, \mathcal{M}_g)$ in (10) by $p(Y | \Lambda, \Sigma^g, \mathcal{M}_g)$ and then obtain an analytic form for $p(Y | \mathcal{M}_g)$ based on Laplace approximation of $\int p(Y | \Lambda, \Sigma^g, \mathcal{M}_g) p(\Lambda) d\Lambda$. This is an instance of integrated nested Laplace approximation for estimating $p(Y | \mathcal{M}_g)$, which has been used extensively in latent Gaussian models (Rue et al., 2009). Using the analytic form of $p(Y | \mathcal{M}_g)$, π_g is obtained using (9); see (45) for an analytic form of $\log \pi_g$ when $p(\mathcal{M}_g) = 1/G$.

We estimate π_g s using (45) after estimating Λ^g s and Σ^g s along the ρ - δ grid. Instead of estimating Λ and Σ at each grid index separately, we carefully choose warm starts that lead to efficient estimation of Λ^g s and Σ^g s as we move across the grid. We first estimate Λ for the largest ρ and smallest δ using (8), which is overly dense and is equivalent to the lasso solution in variable selection. It serves as the warm start for estimating Λ^g s at grid indices starting from the largest to the smallest value of ρ while keeping δ fixed. As ρ decreases, the sequence of Λ estimates refine the dense lasso solution further by thresholding its non-zero values to zero. For the next δ , the sequence of Λ estimates for decreasing ρ s change the warm start to Λ estimated at the previous δ and largest ρ . This process is repeated to estimate Λ s for larger δ s, where as Σ is estimated at every grid index using (6). The idea of using the lasso solution as a warm start has been previously employed in non-concave optimization, including variable selection (Candes et al., 2008; Zhang, 2010b; Mazumder et al., 2011).

The loadings matrix estimate $\bar{\Lambda}^{\text{lla}}$ (9) is more stable than its competitors that select optimal Λ based on

a model selection criterion, say BIC (Hirose and Yamamoto, 2014). The finite grid size of tuning parameters limits the possible Λ and Σ estimates represented by the solution surface. This in turn limits the represented BIC values, which depend on the estimates of Λ and Σ . A typical approach is to choose $g^* \in \{1, \dots, G\}$ with the optimum BIC and choose Λ^{g^*} and Σ^{g^*} as optimal estimates. These estimates are sensitive because it is possible that nearby grid indices have very similar BIC values. It indicates that multiple Λ and Σ estimates could perform equally well. This scenario holds for the two real data analyses presented later. On the contrary, $\bar{\Lambda}^{\text{lla}}$ numerically integrates $\Lambda_{p,\delta}$ across the two-dimensional grid of tuning parameters. This limits sensitivity to grid points because numerical integration is accurate in two-dimensions (Rue et al., 2009). We can also construct credible intervals for $\bar{\Lambda}^{\text{lla}}$ using asymptotic covariance matrix and posterior model weights. Finally, if there exists a true model \mathcal{M}_{g^*} , then its posterior model weight π_{g^*} converges to 1 in probability; see Theorem 4.3 for details.

4 Theoretical properties

4.1 Convergence of the parameter updates

We use an Expectation-Maximization algorithm to estimate Σ (6) and Λ (8), which implies that $\Lambda^{\text{lla}^{(t)}}$ and $\Sigma^{\text{lla}^{(t)}}$ converge to their fixed points $\Lambda^{\text{lla}^{(\infty)}} \equiv \Lambda_N^{\text{lla}}$ and $\Sigma^{\text{lla}^{(\infty)}} \equiv \Sigma_N^{\text{lla}}$ under mild assumptions. If we define the parameter vector $\theta = \{\text{vec}(\Lambda^T), \text{diag}(\Sigma)\}$, where vec operates on Λ^T by stacking its columns into a vector, then updates (6) and (8) define the map $\theta^{(t)} \mapsto \theta^{(t+1)}$. The following theorem proves that iterations (6) and (8) retain the monotone ascent property of Expectation-Maximization.

Theorem 4.1 *If $\mathcal{L}(\theta)$ represents the objective function of expandable factor analysis, then $\mathcal{L}(\theta)$ does not decrease at every iteration. Under standard regularity conditions, the sequence $\{\theta^{(t)}\}_{t=1}^{\infty}$ converges to its stationary point $\theta^{(\infty)} \equiv \theta_N^{\text{lla}}$.*

Proof See Appendix C.1. ■

Remark 4.1 *We use the maximum likelihood estimates $\Lambda^{(0)}$ and $\Sigma^{(0)}$ as \sqrt{N} -consistent estimates based on the theoretical results of Bai and Li (2012) in high-dimensional factor analysis. As recommended in practice, the data analyses presented later use multiple refinements in one-step estimation until the value of the objective stabilizes.*

The objective of expandable factor analysis is bi-convex after local linear approximation, so $\theta^{(\infty)}$ may not be a global optimum. The consistency properties for the sequence $\theta^{(t)}$ are improved by using a \sqrt{N} -consistent starting point $\Lambda^{(0)}$ followed by one-step estimation. This holds because the estimation of θ is a special case of M-estimation (Van der Vaart, 2000).

4.2 Asymptotic normality and consistency of the parameter estimates

Following Zou (2006) and Zou and Li (2008), we show that the estimates Λ_N^{lla} and Σ_N^{lla} are asymptotically normal and consistent. We impose the following assumptions on the sampling model (1) and the hyperpa-

rameters of the multiscale generalized double Pareto prior:

(A1) $y_n = \Lambda^* z_n + e_n$, where z_n and e_n are independent and identically distributed as $\mathcal{N}_K(0, I_K)$ and $\mathcal{N}_P(0, \Sigma^*)$, respectively, with $\Sigma^* = \text{diag}(\sigma_{11}^{2*}, \dots, \sigma_{PP}^{2*})$.

(A2) If $N \rightarrow \infty$, then $\alpha_{k_N} \rightarrow \infty$, $\alpha_{k_N}/\sqrt{N} \rightarrow 0$, and $\sqrt{N} \eta_{k_N} \rightarrow c_k > 0$ for $k = 1, \dots, K$.

Remark 4.2 Assumption (A1) implies that y_n follows the sampling model (1) for every n .

Remark 4.3 The sufficient conditions of Lemma 2.2 are extended to satisfy Assumption (A2) as follows:

- (I) $\alpha_{k_N} = \delta^k N^{\frac{\gamma}{2}}$ or $\alpha_{k_N} = \delta^k \log N$ and $\eta_{k_N} = \rho/\sqrt{N}$ for $k = 1, \dots, K$ and $0 < \gamma < 1$;
- (II) $\alpha_{k_N} = \delta N^{\frac{\gamma}{2}}$ or $\alpha_{k_N} = \delta \log N$ and $\eta_{k_N} = \rho^k/\sqrt{N}$ for $k = 1, \dots, K$ and $0 < \gamma < 1$;
- (III) $\alpha_{k_N} = \delta^k N^{\frac{\gamma}{2}}$ or $\alpha_{k_N} = \delta^k \log N$ and $\eta_{k_N} = \rho^k/\sqrt{N}$ for $k = 1, \dots, K$ and $0 < \gamma < 1$.

In what follows, set $\mathcal{A} \subseteq \{1, \dots, P\}$, and if $v \in \mathbb{R}^P$, then $v_{\mathcal{A}}$ represents v restricted to elements in \mathcal{A} .

Theorem 4.2 Assume that (A1)–(A2) hold, Λ_N^{lla} and Σ_N^{lla} represent the fixed points of maximum a posteriori estimation, Λ^* and Σ^* are the true loadings and residual variance matrices, and $\mathcal{A}_{p_N} = \{k \mid \lambda_{pk}^{\text{lla}} \neq 0 \text{ for } k = 1, \dots, K\}$ for $p = 1, \dots, P$. Then, Λ_N^{lla}

- (i) consistently estimates non-zero loadings: $\lim_{N \uparrow \infty} \mathbb{P}\{\mathcal{A}_{p_N} = \mathcal{A}_p^* \forall p = 1, \dots, P\} = 1$;
- (ii) asymptotic normality holds for non-zero loadings: $\sqrt{N}(\Lambda_{p_{\mathcal{A}_{p_N}}}^{\text{lla}} - \Lambda_{p_{\mathcal{A}_{p_N}}}^*) \rightarrow \mathcal{N}(0, I_{|\mathcal{A}_{p_N}|})$ in distribution for $p = 1, \dots, P$;

and Σ_N^{lla} is asymptotically normal and consistently estimates Σ^* .

Proof See Appendices C.2 and C.3. ■

Remark 4.4 The estimate of Λ has two main advantages. First, a Laplace approximation can be used to obtain standard errors of parameter estimates across a grid of tuning parameters. Second, any Λ_N^{lla} can be used as a warm start for sampling from the posterior distribution of Λ , which avoids exploration of the large model space. These samples can be used to construct posterior credible intervals after accounting for model uncertainty.

Remark 4.5 While the proof of Theorem 4.2 does not rely on any identifiability conditions on Λ^* , our data analyses use the popular lower triangular form of the loadings matrix for comparisons (Bai and Li, 2012).

4.3 Model selection consistency

Expandable factor analysis assigns posterior probability 1 to the true model when N is large. Assume that the ρ - δ grid is fine enough so that grid index g^* corresponds to the true factor analysis model \mathcal{M}_{g^*} , then expandable factor analysis is model selection consistent.

Theorem 4.3 *If Assumptions (A1) and (A2) of Theorem 4.2 hold, the eigenvalues of Ψ^{g^*} are bounded away from zero, and the overfitted factor analysis model for any grid index g satisfies $\mathcal{A}_{p_N}^{g^*} \subset \mathcal{A}_{p_N}^g$ for $p = 1, \dots, P$, then $\lim_{N \uparrow \infty} \mathbb{P}\{\pi_{g^*} = 1\} = 1$ and $\lim_{N \uparrow \infty} \mathbb{P}\{\bar{\Lambda}^{\text{lla}} = \Lambda^*\} = 1$.*

Proof See Appendix C.5. ■

5 Data Analysis

5.1 Notation and setup

We applied Hirose & Yamamoto’s method, sparse principal components (Witten et al., 2009), and expandable factor analysis to simulated and real data. We used the multiscale generalized double Pareto prior that satisfied sufficient condition (I) in Remark 4.3 and set $\alpha_k = \delta^k \log N$ and $\eta_k = \rho/\sqrt{N}$. We compared the performance of the three methods using cumulative number of non-zero loadings across factors, cumulative proportion of explained variance across factors, and root mean square error in estimating the loadings matrix. These metrics are referred to as *CNNL*, *CPEV*, and *RMSE* throughout this section. If Λ was the estimated loadings matrix, then *CNNL* for factor k , *CPEV* for factor k , and *RMSE* were defined as follows:

$$\text{CNNL}_k = \sum_{l=1}^k \sum_{p=1}^P 1_{\Lambda_{pl} \neq 0}, \text{CPEV}_k = \frac{\text{tr}(\lambda_{1:k} \lambda_{1:k}^T)}{\text{tr}(\sum_{n=1}^N \mathbf{y}_n \mathbf{y}_n^T)/N}, \text{RMSE} = \sqrt{\sum_{k=1}^K \sum_{p=1}^P \frac{(\lambda_{pk}^* - \lambda_{pk})^2}{KP}}, \quad (11)$$

where $1_{\Lambda_{pl} \neq 0}$ was 1 for non-zero Λ_{pl} and 0 otherwise, $\lambda_{1:k}$ represented a matrix containing the first k columns of Λ , and λ_{pk}^* was the true value of λ_{pk} . We assumed a *lower triangular form* for Λ^* . In Hirose & Yamamoto’s method and sparse principal components, the number of factors in Λ was chosen using their recommended approaches and the estimated Λ was rotated to a lower triangular form for comparisons using *CNNL* and *RMSE*; *CPEV* _{k} was invariant to rotations of Λ .

5.2 Application of expandable factor analysis to simulated data

We simulated lower-triangular loadings matrices for the following signal-to-noise ratio scenarios:

- (a) *Sparse-High*: sparse loadings matrix with high signal-to-noise ratio;
- (b) *Dense-High*: dense loadings matrix with high signal-to-noise ratio;
- (c) *Dense-Low*: dense loadings matrix with low signal-to-noise ratio.

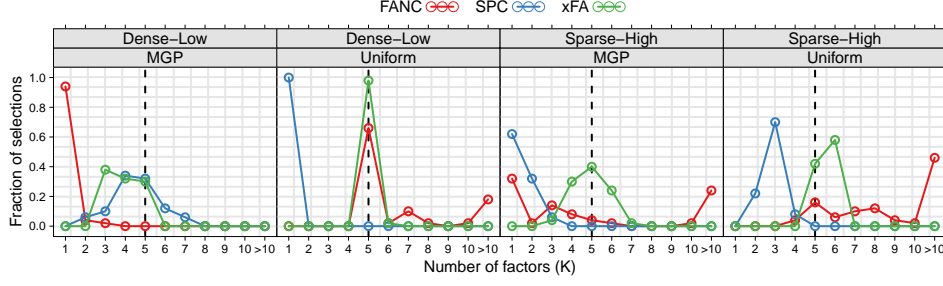


Figure 1: Fraction of selected factors across fifty simulation replications. Panels represent the four simulation settings, x-axis represents the number of estimated factors, and y-axis shows the fraction of times a factor is selected. The true number of factors is 5 (black dashed line). MGP, multiplicative gamma process; FANC, Hirose & Yamamoto’s method; SPC, sparse principal components; xFA, expandable factor analysis.

The loadings in scenarios (a) – (c) were simulated from two prior distributions: (i) multiplicative gamma process and (ii) uniform. The magnitude of the loadings simulated from multiplicative gamma process prior decreased across factors, a pattern similar to the proposed prior, whereas loadings simulated from a uniform prior were biased against this pattern. We simulated Λ following the settings of Bhattacharya and Dunson (2011). Specifically, sparsity was imposed in Λ by having only 10 non-zero loadings in each column. The non-zero loadings in the k th column of Λ were in rows k to $(k+9)$. For each of the six simulation scenarios, $P = 500$, $K = 5$, and $N = \lceil P \log P \rceil$. The error variances were sampled from a uniform prior. Using (1), we generated 50 data sets for each of the six simulation scenarios. We applied the three methods to each of the simulated data sets by fixing the maximum number of factors at 20. The simulation results were compared using CNL, CPEV, and RMSE; see (11).

The three methods performed equally well across the three metrics for the Dense-High simulations using both the priors; the results are not presented here. The performance of the three methods differed substantially for the Sparse-High and Dense-Low simulations (Figure 1). Expandable factor analysis performed well in estimating the true number of factors across four simulation settings, especially outperforming its competitors for Sparse-High data. Sparse principal components under-performed mainly because its greedy estimation algorithm underestimated the rank- K structure of the covariance matrix, which was especially noticeable in Sparse-High and in Dense-Low data. Hirose & Yamamoto’s method for factor selection was unstable across all simulation settings; in some cases, it even selected the maximum allowed 20 factors. A similar behavior was observed for this method in the two real data analyses presented later.

All the three methods underestimated CNL (11) for Dense-Low data and overestimated CNL for Sparse-High data (Figure 2). Expandable factor analysis, however, was closest to the true value with minimum variance across all factors and across all simulation settings; performance can be improved further by setting loadings equal to zero when 95% asymptotic credible intervals include zero; see (43). Sparse principal components under-performed mainly due to its upper bound of \sqrt{P} on the ℓ_1 norm of individual columns of Λ . The upper-bound underestimated the true CNL for Dense-Low data (Figure 2a) and overestimated the true CNL for Sparse-High data (Figure 2b). The factor analytic model of Hirose & Yamamoto’s method led to its improved performance for estimating CNL compared to that in estimating K .

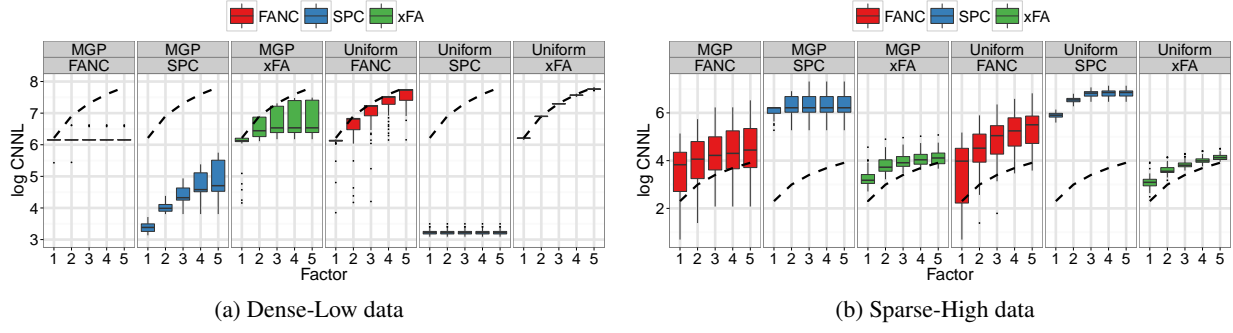


Figure 2: CNNL across factors. The first and last three panels respectively show results for multiplicative gamma process and uniform priors, and y-axis shows log CNNL (11). All panels have superimposed true log CNNL curve (black dashed line). MGP, multiplicative gamma process; FANC, Hirose & Yamamoto’s method; SPC, sparse principal components; xFA, expandable factor analysis; CNNL, cumulative number of non-zero loadings.

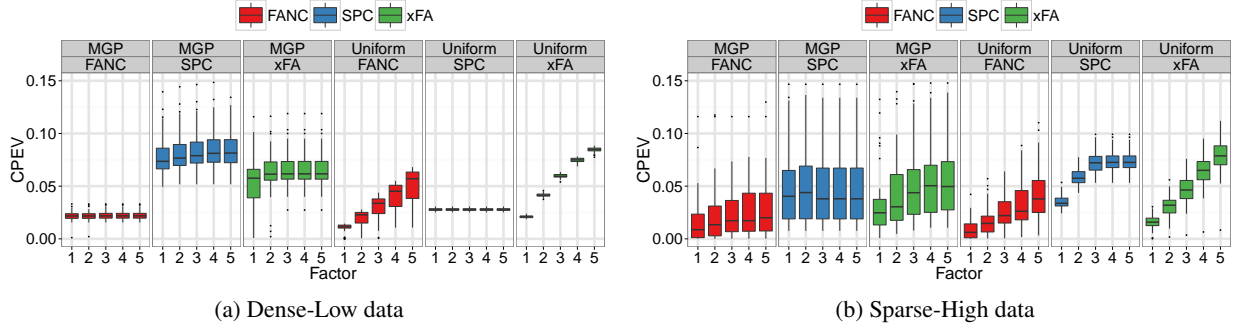


Figure 3: CPEV across factors. The first and last three panels respectively show results for MGP and uniform priors, and y-axis shows CPEV (11). MGP, multiplicative gamma process; FANC, Hirose & Yamamoto’s method; SPC, sparse principal components; xFA, expandable factor analysis; CPEV, cumulative proportion of explained variance.

Expandable factor analysis performed better than the two methods in terms of CPEV in settings where it did not underestimate K (Figure 3); the better performance was noticeable for Dense-Low and Sparse-High data simulated using a uniform prior, which were biased against its sampling model. The greedy approximation of covariance matrix ensured that sparse principal components explained a significant proportion of variance, but it failed to capture the factor analytic structure. Hirose & Yamamoto’s method under-performed due to overestimation of K .

Expandable factor analysis showed better estimation accuracy for Sparse-High data, whereas all the three methods had comparable estimation accuracy for Dense-Low data (Figure 4). Hirose & Yamamoto’s method and expandable factor analysis mostly underestimated the magnitude of loadings simulated from the multiplicative gamma process prior in Dense-Low data. This happens because the loadings simulated from this prior decay more quickly than modeled by Hirose & Yamamoto’s method or expandable factor analysis; see also Figure 1.

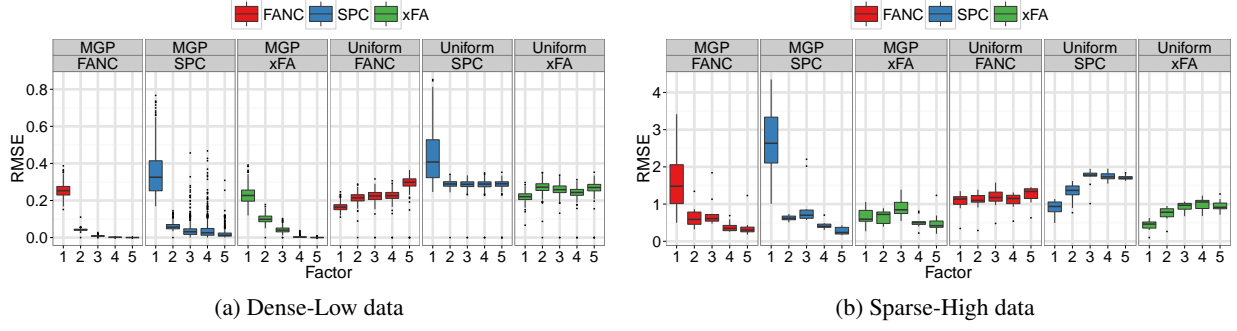


Figure 4: RMSE in estimation of non-zero loadings. The first and last three panels respectively show results for MGP and uniform priors, and y-axis shows RMSE (11). MGP, multiplicative gamma process; FANC, Hirose & Yamamoto’s method; SPC, sparse principal components; xFA, expandable factor analysis; RMSE, root mean square error.

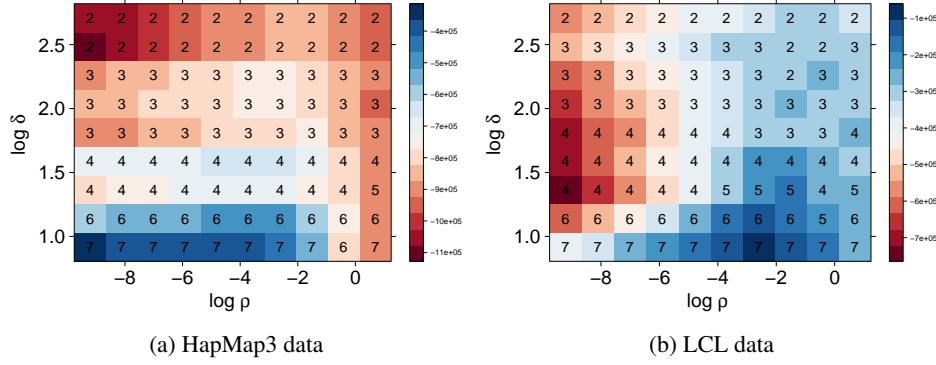


Figure 5: Surface of log model weights $\log \pi_g$ (45). The heatmaps represent $\log \pi_g$ as g varies across $\log \rho$ and $\log \delta$ grid. The number in each cell across the grid represents the number of non-zero columns in the estimated loadings matrix at the corresponding value of ρ and δ .

The more precise and stable performance of expandable factor analysis over its competitors was due in part to Bayesian model averaging. Expandable factor analysis combined results from several good models across the ρ - δ grid and removed sensitivity to hyperparameters and selection of K .

5.3 Application of expandable factor analysis to microarray data

We analyzed two publicly available microarray data sets: *HapMap3* data with gene expression measurements in 608 samples and 17,229 genes (Thorisson et al., 2005) and *LCL* data with gene expression measurements in human immortalized cell lines with 480 samples and 8,718 genes (Brown et al., 2013). We partitioned the samples using 5-fold cross-validation, applied the three methods with maximum number of factors fixed at 20, and compared their performances using CNL and CPEV (11). The posterior model weights π_g (45) were estimated by varying $\log \rho$ and $\log \delta$ on a grid (Figures 5a and 5b). The results across all 5 folds were stable and similar, so only overall results based on the median are presented.

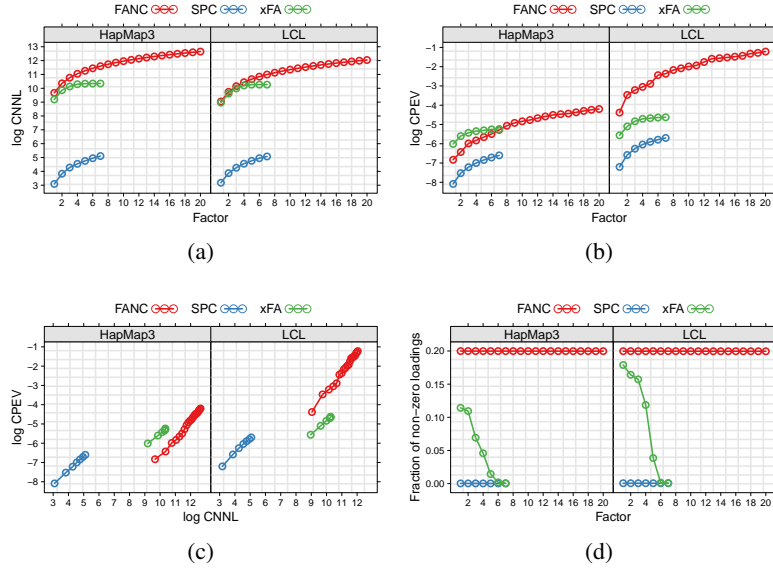


Figure 6: (a) CNL (11) across factors; (b) CPEV (11) across factors; (c) CPEV as a function of CNL; (d) fraction of non-zero loadings across factors. The first and second panels respectively represent HapMap3 and LCL data for (a) – (d). FANC, Hirose & Yamamoto’s method; SPC, sparse principal components; xFA, expandable factor analysis; CNL, cumulative number of non-zero loadings; CPEV, cumulative proportion of explained variance.

The $\log \pi_g$ surfaces for HapMap3 and LCL data were multi-modal with nearby values being fairly similar (Figures 5a and 5b). Model averaging protected us from solving the difficult optimization problem of finding the optimal tuning parameters ρ and δ from the multi-modal surface of $\log \pi_g$. The model averaged estimate of Λ selected $K = 7$ out of the maximum 20 factors for HapMap3 and LCL data across all 5 folds. Hirose & Yamamoto’s method and sparse principal components respectively selected $K = 20$ and $K = 1$ out of the maximum 20 factors for both real data across all 5 folds. Furthermore, the selected tuning parameters for Hirose & Yamamoto’s method were on the boundaries of their respective parameter spaces across all 5 folds. Results for the two methods did not improve despite using several settings of tuning parameters. We applied sparse principal components after fixing $K = 7$ to compare its results with expandable factor analysis.

The selection of tuning parameters on the boundary for Hirose & Yamamoto’s method could be explained from multi-modal $\log \pi_g$ surface (Figure 5). Both BIC and $\log \pi_g$ are based on Laplace approximation and define posterior weights for models depending on their two tuning parameters. Hirose & Yamamoto’s method selected tuning parameters that correspond to a model with high π_g on the equivalent ρ - δ grid. The multi-modal $\log \pi_g$ surface implies that the BIC surface might be multi-modal; therefore, optimal tuning parameter selection over the multi-modal BIC surface could lead to solutions on the boundary.

Hirose & Yamamoto’s method selected tuning parameters on the boundary, so the estimated Λ was overly dense in HapMap3 data and incompletely regularized in LCL data. This also explained its high CNL for both real data (Figure 6a). On the contrary, Λ estimates for both real data in sparse principal components were overly sparse due to the upper bound of \sqrt{P} on the ℓ_1 norm of individual columns of Λ .

Its CPEV was the smallest among the three methods for both real data. The CPEV for expandable factor analysis was largest across all its factors for HapMap3 data, where as CPEV for Hirose & Yamamoto's method was the largest across all factors for LCL data (Figure 6b). Defining a good fit as high values of CPEV for low values of CNL implies that expandable factor analysis fit HapMap3 data better than its competitors; that is, sparse Λ explained a large fraction of variance. However, it was not clear if Hirose & Yamamoto's method was a good fit for LCL data because its higher CPEV than its competitors came at the cost of high CNL (Figure 6c).

The results for expandable factor analysis in both real data agreed with its theory. The fraction of non-zero loadings across factors was smallest for sparse principal components due to its overly sparse Λ , where as this metric was constant and largest for Hirose & Yamamoto's method due to its overly dense Λ . Expandable factor analysis used a structured prior for regularization that enabled it to have better control over the fraction of non-zero loadings across factors. Specifically, its fraction of non-zero loadings in the first factor was very close to that of Hirose & Yamamoto's method; and this metric in the last factor was very close to that of sparse principal components. Between the two extremes, the fraction of non-zero loadings decayed exponentially across factors (Figure 6d); therefore, the multiscale generalized double Pareto prior satisfied property (d).

Acknowledgement

Srivastava and Dunson were partially funded by a grant from the National Institute of Environmental Health Sciences of the National Institutes of Health, with Srivastava receiving additional funding from the Statistical and Applied Mathematical Sciences Institute. Engelhardt was partially funded through grants from the National Institutes of Health. Thanks are given to David Lawlor, Minh Pham and Cheng Li for helpful conversations.

A Properties of the multiscale generalized double Pareto prior

A.1 Proof of Lemma 2.1

We prove that $\mathbb{P}_{\text{load}}\{\mathcal{C}_{\text{load}}\} = 1$ using the definition of $\mathcal{C}_{\text{load}}$. Specifically,

$$\begin{aligned} \mathbb{P}_{\text{load}}\{\mathcal{C}_{\text{load}}\} &= \mathbb{P}_{\text{load}}\{\Lambda \mid \max_{1 \leq p \leq P} \sum_{k=1}^{\infty} \lambda_{pk}^2 < \infty\} = 1 - \lim_{t \uparrow \infty} \mathbb{P}_{\text{load}}\{\Lambda \mid \max_{1 \leq p \leq P} \sum_{k=1}^{\infty} \lambda_{pk}^2 \geq t\} \\ &\geq 1 - \lim_{t \uparrow \infty} \sum_{p=1}^P \mathbb{P}_{\text{load}}\{\Lambda \mid \sum_{k=1}^{\infty} \lambda_{pk}^2 \geq t\} \geq 1 - P \lim_{t \uparrow \infty} \frac{\sum_{k=1}^{\infty} \mathbb{E}[\lambda_{1k}^2]}{t} = 1 - P \lim_{t \uparrow \infty} \frac{\sum_{k=1}^{\infty} V[\lambda_{1k}]}{t}. \end{aligned} \quad (12)$$

Armagan et al. (2013) show that $\lambda_{1k} \sim \text{GDP}(\alpha, \eta)$ has $V[\lambda_{1k}] = 2\eta^2(\alpha - 1)^{-1}(\alpha - 2)^{-1}$ for $\alpha > 2$, so

$$\begin{aligned} \sum_{k=1}^{\infty} V[\lambda_{1k}] &= 2 \sum_{k=1}^{\infty} \eta_k^2 \frac{1}{\alpha_k - 1} \frac{1}{\alpha_k - 2} \leq 2 \sum_{k=1}^{\infty} \eta_k^2 \frac{1}{\alpha_k - 2} \frac{1}{\alpha_k - 2} \leq 2 \sum_{k=1}^{\infty} \frac{\eta_k^2}{\alpha_k^2} \left(1 - \frac{2}{\alpha_k}\right)^{-2} \\ &= 2 \sum_{k=1}^{\infty} \frac{\eta_k^2}{\alpha_k^2} \left(1 + \frac{4}{\alpha_k} + o\left(\frac{1}{\alpha_k}\right)\right) < 2(1 + 2 + \mathcal{O}(1)) \sum_{k=1}^{\infty} \frac{\eta_k^2}{\alpha_k^2} < \infty \end{aligned} \quad (13)$$

if $\alpha_k > 2$ and $\eta_k/\alpha_k = \mathcal{O}(1/k^m)$ for $m > 0.5$; therefore, $\sum_{k=1}^{\infty} V[\lambda_{1k}]$ in (12) is bounded and $\mathbb{P}_{\text{load}}\{\mathcal{C}_{\text{load}}\} = 1$.

A.2 Proof of Lemma 2.2

Expandable factor analysis finds $K_0 = K(P, \delta, \rho, \epsilon)$ that upper bounds $\mathbb{P}\{d_{\infty}(\Omega, \Omega^{K_0}) \geq \epsilon\}$ by ϵ .

$$\begin{aligned} \mathbb{P}\{\Omega^{K_0} \mid d_{\infty}(\Omega, \Omega^{K_0}) < \epsilon\} &= \mathbb{P}\{\Omega^{K_0} \mid \max_{1 \leq i, j \leq P} |\Omega_{ij} - \Omega_{ij}^{K_0}| < \epsilon\} \\ &= 1 - \mathbb{P}\{\Omega^{K_0} \mid \max_{1 \leq i, j \leq P} \sum_{k=K_0+1}^{\infty} \lambda_{ik} \lambda_{jk} \geq \epsilon\} \\ &\geq 1 - \sum_{i=1}^P \sum_{j=1}^P \mathbb{P}\{\Omega^{K_0} \mid \sum_{k=K_0+1}^{\infty} \lambda_{ik} \lambda_{jk} \geq \epsilon\} \\ &\geq 1 - \sum_{i=1}^P \sum_{j=1}^P \mathbb{P}\{\Omega^{K_0} \mid \sum_{k=K_0+1}^{\infty} |\lambda_{ik} \lambda_{jk}| \geq \epsilon\} \\ &\geq 1 - \sum_{k=K_0+1}^{\infty} \frac{\mathbb{E}[\sum_{i=1}^P \sum_{j=1}^P |\lambda_{ik}| |\lambda_{jk}|]}{\epsilon}. \end{aligned}$$

Using Hölder's inequality and noticing that λ_{ik} 's are sampled independently from $\text{GDP}(\alpha_k, \eta_k)$,

$$\begin{aligned} \mathbb{P}\{\Omega^{K_0} \mid d_{\infty}(\Omega, \Omega^{K_0}) < \epsilon\} &\geq 1 - \sum_{k=K_0+1}^{\infty} \frac{\mathbb{E}[(\sum_{i=1}^P |\lambda_{ik}|)^2]}{\epsilon} \\ &= 1 - \frac{P^2}{\epsilon} \sum_{k=K_0+1}^{\infty} (V[\lambda_{1k}] + \mathbb{E}^2[\lambda_{1k}]). \end{aligned} \quad (14)$$

The last summation in (14) depends on the sufficient conditions of Lemma 2.2:

(I) When $\alpha_k = \delta^k$ and $\eta_k = \rho$, where $\delta > 2$ and $\rho > 0$, then

$$\begin{aligned} \sum_{k=K_0+1}^{\infty} \mathbb{E}^2[\lambda_{1k}] &= \sum_{k=K_0+1}^{\infty} \frac{\eta_k^2}{\alpha_k^2} = \frac{\rho^2}{\delta^{2K_0+2}} \sum_{k=0}^{\infty} \frac{1}{\delta^{2k}} = \mathcal{O}\left(\frac{1}{\delta^{2K_0}}\right); \\ \sum_{k=K_0+1}^{\infty} V[\lambda_{1k}] &\leq 2\rho^2 \sum_{k=K_0+1}^{\infty} \frac{1}{\delta^{2k}} \left(1 - \frac{2}{\delta^{2k}}\right)^{-2} = \mathcal{O}\left(\frac{1}{\delta^{2K_0}}\right). \end{aligned}$$

(II) When $\alpha_k = \delta$ and $\eta_k = \rho^k$, where $\delta > 2$ and $1 > \rho > 0$, then

$$\begin{aligned} \sum_{k=K_0+1}^{\infty} \mathbb{E}^2[|\lambda_{1k}|] &= \sum_{k=K_0+1}^{\infty} \frac{\eta_k^2}{\alpha_k^2} = \frac{\rho^{2K_0+2}}{\delta^2} \sum_{k=0}^{\infty} \rho^{2k} = \mathcal{O}(\rho^{2K_0}); \\ \sum_{k=K_0+1}^{\infty} V[\lambda_{1k}] &= \frac{2}{(\delta-1)(\delta-2)} \sum_{k=K_0+1}^{\infty} \rho^{2k} = \mathcal{O}(\rho^{2K_0}). \end{aligned}$$

(III) When $\alpha_k = \delta^k$ and $\eta_k = \rho^k$, where $\delta > 2$ and $\delta > \rho > 0$, then

$$\begin{aligned} \sum_{k=K_0+1}^{\infty} \mathbb{E}^2[|\lambda_{1k}|] &= \sum_{k=K_0+1}^{\infty} \frac{\eta_k^2}{\alpha_k^2} = \frac{\rho^{2K_0+2}}{\delta^{2K_0+2}} \sum_{k=0}^{\infty} \left(\frac{\rho}{\delta}\right)^{2k} = \mathcal{O}\left(\frac{\rho^{2K_0}}{\delta^{2K_0}}\right); \\ \sum_{k=K_0+1}^{\infty} V[\lambda_{1k}] &\leq 2\rho^2 \sum_{k=K_0+1}^{\infty} \frac{\rho^{2k}}{\delta^{2k}} \left(1 - \frac{2}{\delta^{2k}}\right)^{-2} = \mathcal{O}\left(\frac{\rho^{2K_0}}{\delta^{2K_0}}\right). \end{aligned}$$

Using (14), K_0 that satisfies $\mathbb{P}\{\Omega^{K_0} \mid d_{\infty}(\Omega, \Omega^{K_0}) < \epsilon\} \geq 1 - \epsilon$ depends as follows on the sufficient conditions of Lemma 2.2:

- (I) $\frac{P^2}{\epsilon} \mathcal{O}\left(\frac{1}{\delta^{2K_0}}\right) \leq \epsilon \implies K_0 = \mathcal{O}(\log^{-1} \delta \log \frac{P}{\epsilon});$
- (II) $\frac{P^2}{\epsilon} \mathcal{O}(\rho^{2K_0}) \leq \epsilon \implies K_0 = \mathcal{O}\left(\log^{-1} \frac{1}{\rho} \log \frac{P}{\epsilon}\right);$
- (III) $\frac{P^2}{\epsilon} \mathcal{O}\left(\frac{\rho^{2K_0}}{\delta^{2K_0}}\right) \leq \epsilon \implies K_0 = \mathcal{O}\left(\log^{-1} \frac{\delta}{\rho} \log \frac{P}{\epsilon}\right).$

B Computations in expandable factor analysis

B.1 Log-posterior based on the Expectation-Maximization algorithm and local linear approximation

Denoting Z as *missing* data and Y as *observed* data, expandable factor analysis extends the maximum likelihood estimation approach of Rubin and Thayer (1982) to maximum a posteriori estimation of Λ and Σ using multiscale generalized double Pareto prior on Λ and Jeffreys' prior on the diagonal elements of Σ . This involves maximization of $\mathbb{E}\{\log p(Z, \Lambda, \Sigma \mid Y, \Lambda^{(t)}, \Sigma^{(t)}, \alpha_{1:K}, \eta_{1:K})\}$

$$\begin{aligned} &= - \sum_{p=1}^P \frac{N}{2} \frac{(S_{yy})_{pp} + (\Lambda \mathbb{E}[S_{zz} \mid Y, \Lambda^{(t)}, \Sigma^{(t)}] \Lambda^T)_{pp} - 2(\mathbb{E}[S_{yz} \mid Y, \Lambda^{(t)}, \Sigma^{(t)}] \Lambda^T)_{pp}}{\sigma_{pp}^2} \\ &\quad - \sum_{p=1}^P \sum_{k=1}^K (\alpha_k + 1) \log \left(1 + \frac{|\lambda_{pk}|}{\eta_k}\right) - \frac{N+2}{2} \sum_{p=1}^P \log \sigma_{pp}^2 \\ &\equiv - \sum_{p=1}^P \log p_{\text{mis}}(\lambda_p, \sigma_{pp}^2 \mid S_{yy}, \Psi^{(t)}, \hat{\Lambda}^{(t)}) - \frac{N+2}{2} \sum_{p=1}^P \log \sigma_{pp}^2, \end{aligned} \tag{15}$$

where the analytic forms of conditional expectations, $\Psi^{(t)}$, and $\hat{\Lambda}^{(t)}$ are as follows:

$$\begin{aligned} S_{yy} &= \frac{1}{N} \sum_{n=1}^N y_n y_n^T; \quad S_{zz} = \frac{1}{N} \sum_{n=1}^N z_n z_n^T; \quad S_{yz} = \frac{1}{N} \sum_{n=1}^N y_n z_n^T; \\ \Delta &= I - \Lambda^T (\Lambda \Lambda^T + \Sigma)^{-1} \Lambda; \quad \Gamma = (\Lambda \Lambda^T + \Sigma)^{-1} \Lambda; \quad \Psi = \Delta + \Gamma^T S_{yy} \Gamma; \quad \hat{\Lambda} = S_{yy} \Gamma; \\ E[S_{zz} | Y, \Lambda^{(t)}, \Sigma^{(t)}] &= \Delta^{(t)} + \Gamma^{(t)T} S_{yy} \Gamma^{(t)} = \Psi^{(t)}; \quad E[S_{yz} | Y, \Lambda^{(t)}, \Sigma^{(t)}] = \hat{\Lambda}^{(t)}. \end{aligned} \quad (16)$$

We observe that (15) splits into P separate terms corresponding to each dimension of Y ; therefore, estimating Λ and Σ by maximizing (15) at the $(t+1)$ th iteration is equivalent to separately minimizing P objectives of the form

$$\log p_{\text{mis}}(\lambda_p, \sigma_{pp}^2 | S_{yy}, \Psi^{(t)}, \hat{\Lambda}^{(t)}) + \frac{N+2}{2} \log \sigma_{pp}^2, \quad (17)$$

with respect to λ_p and σ_{pp}^2 for $p = 1, \dots, P$; each of which is bi-convex. Following the standard Expectation-Maximization approach, for $p = 1, \dots, P$, we first fix σ_{pp}^2 at $\sigma_{pp}^{2(t)}$ and minimize $\log p_{\text{mis}}(\lambda_p, \sigma_{pp}^{2(t)} | S_{yy}, \Psi^{(t)}, \hat{\Lambda}^{(t)})$ in (17). We then fix λ_p at $\lambda_p^{(t)}$ in (17) and minimize with respect to σ_{pp}^2 , and repeat these steps until convergence to the local optimum.

B.2 Block coordinate descent algorithm for estimation of Λ

We derive the coordinate descent updates for λ_p^{lla} at the $(t+1)$ th iteration given $\lambda_p^{\text{lla}(t)}$. We suppress (t) in w_p and X to ease notation. The cycle of coordinate descent algorithm for updating $\lambda_p^{\text{lla}(t)}$ to $\lambda_p^{\text{lla}(t+1)}$ starts by initializing $\tilde{\Lambda}^{(1)} = \Lambda^{\text{lla}(t)}$ and $(i+1)$ th cycle updates $\tilde{\Lambda}_p^{(i)}$ using (7) as

$$\tilde{\lambda}_{pk}^{(i+1)} = \underset{\lambda_{pk}}{\text{argmin}} \frac{\lambda_{pk}^2 X_k^T X_k + 2\lambda_{pk} (\tilde{\Lambda}_{p,(-k)}^{(i)})^T X_{(-k)}^T X_k - X_k^T w_p}{2} + \frac{(\alpha_k + 1) \sigma_{pp}^{2(t)}}{(\eta_k + |\lambda_{pk}^{(t)}|)N} |\lambda_{pk}|.$$

The solution of this convex program is obtained as a simple extension of `glmnet` updates, so that

$$\tilde{\lambda}_{pk}^{(i+1)} = \frac{\text{sign}(l_{pk}^{(i)})}{X_k^T X_k} \left(|l_{pk}^{(i)}| - \frac{(\alpha_k + 1) \sigma_{pp}^{2(t)}}{(\eta_k + |\lambda_{pk}^{(t)}|)N} \right)_+, \quad (18)$$

where $l_{pk}^{(i)} = X_k^T w_p - \tilde{\Lambda}_{p,(-k)}^{(i)} X_{(-k)}^T X_k$. Noticing that $\Psi = X^T X$ yields (8), we also exploit the form of (18) and use it to update the k th column of $\tilde{\Lambda}^{(i)}$. This modification results in K block updates for $\tilde{\Lambda}^{(i)}$ in a single cycle of expandable factor analysis coordinate descent algorithm. These update cycles are repeated multiple times until the change in $\tilde{\Lambda}$ is negligible, and then we set $\Lambda^{\text{lla}(t+1)} = \tilde{\Lambda}^{(\infty)}$.

B.3 Implementation and computational complexity

The time complexity of fitting expandable factor analysis without model averaging equals the cost of performing P parallel penalized regression problems of dimension $K = \mathcal{O}(\log P)$. Expandable factor analysis

uses data in form of S_{yy} and forming such a matrix requires $\mathcal{O}(NP^2)$ cost upfront, which is greatly reduced in practice by exploiting the structure of the data. Further, $K-1$ Householder QRs for constructing $\Lambda^{(0)}$ cost $\mathcal{O}(P \log^2 P)$ (Golub and Van Loan, 2012); therefore, the total cost of constructing $\Lambda^{(0)}$ is $\mathcal{O}(P \log^2 P + NP^2)$. Each expandable factor analysis iteration calculates matrices Γ, Δ, Ψ , and $\hat{\Lambda}$ (16) that involve K -dimensional matrix multiplications and inversions. These K -dimensional matrix computations are much more efficient than those in P dimensions and have a total time complexity of $\mathcal{O}(\log^3 P)$. Using these matrices one penalized regression costs $\mathcal{O}(\log P)$, so the total time complexity of each expandable factor analysis iteration is $\mathcal{O}(P \log P + \log^3 P)$. In most practical applications $\log^2 P \ll P$, therefore the time complexity of T iterations of expandable factor analysis is $\mathcal{O}(TP \log P)$.

Algorithm 1 Expandable Factor Analysis using block coordinate descent

Input $\Lambda^{(0)}, \Sigma^{(0)}, S_{yy}, \alpha_{1:K}, \eta_{1:K}$, and maximum iterations for coordinate descent (`maxiter`) and EM (`xfaiter`) algorithms
While polishing of $\Lambda^{(t)}$ and $\Sigma^{(t)}$ is required & $t < \text{xfaiter}$
Set $\Gamma^{(t)} = (\Lambda^{(t)} \Lambda^{(t)T} + \Sigma^{(t)})^{-1} \Lambda^{(t)}, \Delta^{(t)} = I - \Lambda^{(t)T} \Gamma^{(t)}, \Psi^{(t)} = \Delta^{(t)} + \Gamma^{(t)T} S_{yy} \Gamma^{(t)}$, and $\hat{\Lambda}^{(t)} = S_{yy} \Gamma^{(t)}$
Set $C^{(t)} = \text{Cholesky}(\Psi^{(t)})$, $X^{(t)} = C^{(t)T}$, and $\tilde{\Lambda}^{(0)} = \Lambda^{(t)}$
For $k = 1, \dots, K$
Set $c_k^{(t)} = \frac{\Sigma^{(t)}}{N} \left(\frac{\alpha_k + 1}{\eta_k + |\Lambda_{1k}^{(t)}|}, \dots, \frac{\alpha_k + 1}{\eta_k + |\Lambda_{pk}^{(t)}|}, \dots, \frac{\alpha_k + 1}{\eta_k + |\Lambda_{pk}^{(t)}|} \right)^T$ and $i = 0$
While polishing of $\tilde{\Lambda}_k^{(i)}$ is required & $i < \text{maxiter}$
For $p = 1, \dots, P$
Set $l_{pk}^{(i)} = X_k^T w_p - \tilde{\Lambda}_{p,(-k)}^{(i)} X_{(-k)}^T X_k$
Set $\tilde{\lambda}_{pk}^{(i+1)} = \frac{\text{sign}(l_{pk}^{(i)})}{X_k^T X_k} \left(|l_{pk}^{(i)}| - c_{pk}^{(t)} \right)_+$
End for
Update the k th column of the final result $\tilde{\Lambda}_k = \tilde{\Lambda}_k^{(i)}$
For $p = 1, \dots, P$
Set $\sigma_{pp}^{2'} = \frac{N}{N+2} ((S_{yy})_{pp} + \lambda_p^{(t)T} \Psi^{(t)} \lambda_p^{(t)} - 2\tilde{\lambda}_p^{(t)T} \lambda_p^{(t)})$
End for
End while
End for
Set $t = t + 1$ and update parameters $\Lambda^{(t)} = \tilde{\Lambda}$ and $\Sigma^{(t)} = \text{diag}(\sigma_{11}^{2'}, \dots, \sigma_{pp}^{2'})$
End while
Output $\Lambda^{(\infty)}$ and $\Sigma^{(\infty)}$

C Theoretical properties of expandable factor analysis

C.1 Proof of Theorem 4.1

We use (15) and (16) and define

$$\mathcal{Q}(\theta \mid \theta^{(t)}) = - \sum_{p=1}^P \log p_{\text{mis}}(\lambda_p, \sigma_{pp}^2 \mid S_{yy}, \Psi^{(t)}, \hat{\Lambda}^{(t)}) - \frac{N+2}{2} \sum_{p=1}^P \log \sigma_{pp}^2. \quad (19)$$

The local linear approximation of $\mathcal{Q}(\theta \mid \theta^{(t)})$ (19) is

$$\begin{aligned} \mathcal{Q}_{LLA}(\theta \mid \theta^{(t)}) = & - \sum_{p=1}^P \frac{N}{2} \frac{(S_{yy})_{pp} + (\Lambda \Psi^{(t)} \Lambda^T)_{pp} - 2(\widehat{\Lambda}^{(t)} \Lambda^T)_{pp}}{\sigma_{pp}^2} - \frac{N+2}{2} \sum_{p=1}^P \log \sigma_{pp}^2 \\ & - \sum_{p=1}^P \sum_{k=1}^K (\alpha_k + 1) \left\{ \log \left(1 + \frac{|\lambda_{pk}^{(t)}|}{\eta_k} \right) + \frac{\text{sign}(\lambda_{pk}^{(t)})}{\eta_k + |\lambda_{pk}^{(t)}|} (\lambda_{pk} - \lambda_{pk}^{(t)}) \right\}, \end{aligned} \quad (20)$$

and the sampling model of expandable factor analysis (1) and the multiscale generalized double Pareto prior yield the log posterior

$$\mathcal{L}(\theta) = \mathcal{L}_{ML}(\theta) - \sum_{p=1}^P \sum_{k=1}^K (\alpha_k + 1) \log \left(1 + \frac{|\lambda_{pk}|}{\eta_k} \right) - \frac{N+2}{2} \sum_{p=1}^P \log \sigma_{pp}^2, \quad (21)$$

where $\mathcal{L}_{ML}(\theta)$ is the log likelihood defined in Rubin and Thayer (1982) for the maximum likelihood estimation of θ in (1). Similarly, if we represent $\mathcal{Q}_{ML}(\theta \mid \theta^{(t)})$ as the \mathcal{Q} function in maximum likelihood estimation of θ , then Rubin and Thayer (1982) and (20) imply that

$$\begin{aligned} \mathcal{Q}_{LLA}(\theta \mid \theta^{(t)}) = & \mathcal{Q}_{ML}(\theta \mid \theta^{(t)}) - \sum_{p=1}^P \sum_{k=1}^K (\alpha_k + 1) \left\{ \log \left(1 + \frac{|\lambda_{pk}^{(t)}|}{\eta_k} \right) + \frac{\text{sign}(\lambda_{pk}^{(t)})}{\eta_k + |\lambda_{pk}^{(t)}|} (\lambda_{pk} - \lambda_{pk}^{(t)}) \right\} \\ & - \frac{N+2}{2} \sum_{p=1}^P \log \sigma_{pp}^2. \end{aligned} \quad (22)$$

Theorem 1 of Dempster et al. (1977) shows that $\mathcal{Q}_{ML}(\theta^{(t)} \mid \theta^{(t)}) = \mathcal{L}_{ML}(\theta^{(t)})$, so $\mathcal{Q}(\theta^{(t)} \mid \theta^{(t)}) = \mathcal{L}(\theta^{(t)})$ and

$$\mathcal{Q}_{LLA}(\theta^{(t)} \mid \theta^{(t)}) = \mathcal{L}_{ML}(\theta^{(t)}) - \sum_{p=1}^P \sum_{k=1}^K (\alpha_k + 1) \log \left(1 + \frac{|\lambda_{pk}^{(t)}|}{\eta_k} \right) - \frac{N+2}{2} \sum_{p=1}^P \log \sigma_{pp}^2 = \mathcal{L}(\theta^{(t)}), \quad (23)$$

where the last equality follows from (21). Subtracting (22) from (21) yields

$$\mathcal{L}(\theta) - \mathcal{Q}_{LLA}(\theta \mid \theta^{(t)}) = \mathcal{L}_{ML}(\theta) - \mathcal{Q}_{ML}(\theta \mid \theta^{(t)}) + \sum_{p=1}^P \sum_{k=1}^K (\alpha_k + 1) l_{pk}(\lambda_{pk} \mid \lambda_{pk}^{(t)}), \quad (24)$$

where

$$l_{pk}(\lambda_{pk} \mid \lambda_{pk}^{(t)}) = \log \left(1 + \frac{|\lambda_{pk}^{(t)}|}{\eta_k} \right) + \frac{\text{sign}(\lambda_{pk}^{(t)})}{\eta_k + |\lambda_{pk}^{(t)}|} (\lambda_{pk} - \lambda_{pk}^{(t)}) - \log \left(1 + \frac{|\lambda_{pk}|}{\eta_k} \right); \quad (25)$$

We use the fact that \log is a concave function and its tangent is always above its graph; that is, the tangent majorizes \log function and $l_{pk}(\lambda_{pk} \mid \lambda_{pk}^{(t)}) \geq 0$ for any $|\lambda_{pk}| > 0$. Using Lemma 1 and Theorem 1 of

Dempster et al. (1977) $\mathcal{L}_{\text{ML}}(\theta) - \mathcal{Q}_{\text{ML}}(\theta \mid \theta^{(t)}) \geq 0$, then $\mathcal{L}(\theta) - \mathcal{Q}_{\text{LLA}}(\theta \mid \theta^{(t)}) \geq 0$. If the $(t+1)$ th update of θ , $\theta^{(t+1)}$, maximizes $\mathcal{Q}_{\text{LLA}}(\theta \mid \theta^{(t)})$, then we have that

$$\mathcal{L}(\theta^{(t+1)}) \geq \mathcal{Q}_{\text{LLA}}(\theta^{(t+1)} \mid \theta^{(t)}) \geq \mathcal{Q}_{\text{LLA}}(\theta^{(t)} \mid \theta^{(t)}) = \mathcal{L}(\theta^{(t)}), \quad (26)$$

where the last equality follows from (23). The log posterior of expandable factor analysis $\mathcal{L}(\theta)$ is bounded on the parameter space because it depends on Y only through S_{yy} , so the sequence $\{\mathcal{L}(\theta^{(t)})\}_{t=1}^{\infty}$ converges to some $\mathcal{L}(\theta^{(\infty)})$. This however does not imply that the sequence $\{\theta^{(t)}\}_{t=1}^{\infty}$ converges to $\theta^{(\infty)}$. Proposition 1 of Zou and Li (2008) implies that $\theta^{(t)}$ converges to the stationary point $\theta^{(\infty)}$.

C.2 Proof of Theorem 4.2 (asymptotic normality and consistency of Λ^{lla})

We now show the consistency and asymptotic normality of one-step estimators. The proof of consistency and asymptotic normality of Λ^{lla} does not require identifiable loadings matrix or the sufficient conditions of Lemma 2.2. We hide dependence on N to ease notation. Denoting λ_p^{lla} as the expandable factor analysis estimate of λ_p from local linear approximations using $\lambda_p^{(0)}$ as the \sqrt{N} -consistent sequence of estimators, (7) implies that for $p = 1, \dots, P$

$$\lambda_p^{\text{lla}} = \underset{\lambda_p}{\operatorname{argmin}} \frac{N}{2} \frac{\|w_p^{(0)} - X^{(0)}\lambda_p\|^2}{\sigma_{pp}^{2(0)}} + \sum_{k=1}^K \frac{\alpha_k + 1}{\eta_k + |\lambda_{pk}^{(0)}|} |\lambda_{pk}|, \quad (27)$$

where $w_p^{(0)} = \Psi^{(0)-1/2} \hat{\lambda}_p^{(0)}$, *pseudo* design matrix $X^{(0)} = \Psi^{(0)1/2}$, and $\Sigma^{(0)}$ is obtained from (6) using $\lambda^{(0)}$. We denote the true value of λ_p as λ_p^* and define $u_p = (u_{p1}, \dots, u_{pk}, \dots, u_{pK})^T$ and

$$\mathcal{V}(u_p) = \frac{N}{2} \frac{\|w_p^{(0)} - X^{(0)}(\lambda_p^* + \frac{u_p}{\sqrt{N}})\|^2}{\sigma_{pp}^{2(0)}} + \sum_{k=1}^K \frac{\alpha_k + 1}{\eta_k + |\lambda_{pk}^{(0)}|} |\lambda_{pk}^* + \frac{u_{pk}}{\sqrt{N}}|, \quad (28)$$

where vectors are added component-wise. By substituting $u_{pk} = 0$ for $k = 1, \dots, K$ in (28),

$$\mathcal{V}(0) = \frac{N}{2} \frac{\|w_p^{(0)} - X^{(0)}\lambda_p^*\|^2}{\sigma_{pp}^{2(0)}} + \sum_{k=1}^K \frac{\alpha_k + 1}{\eta_k + |\lambda_{pk}^{(0)}|} |\lambda_{pk}^*|, \quad (29)$$

and

$$\begin{aligned}
\mathcal{V}(\mathbf{u}_p) - \mathcal{V}(0) &= \frac{N}{2\sigma_{pp}^{2(0)}} \{ \|\mathbf{w}_p^{(0)} - \mathbf{X}^{(0)}(\lambda_p^* + \frac{\mathbf{u}_p}{\sqrt{N}})\|^2 - \|\mathbf{w}_p^{(0)} - \mathbf{X}^{(0)}\lambda_p^*\|^2 \} \\
&+ \sum_{k=1}^K \frac{\alpha_k + 1}{\eta_k + |\lambda_{pk}^{(0)}|} (|\lambda_{pk}^* + \frac{\mathbf{u}_{pk}}{\sqrt{N}}| - |\lambda_{pk}^*|) \\
&= \frac{1}{2} \mathbf{u}_p^\top \frac{\Psi^{(0)}}{\sigma_{pp}^{2(0)}} \mathbf{u}_p - \mathbf{u}_p^\top \frac{\sqrt{N}(\hat{\lambda}_p^{(0)} - \Psi^{(0)}\lambda_p^*)}{\sigma_{pp}^{2(0)}} \\
&+ \sum_{k=1}^K \frac{\alpha_k + 1}{\eta_k + |\lambda_{pk}^{(0)}|} (|\lambda_{pk}^* + \frac{\mathbf{u}_{pk}}{\sqrt{N}}| - |\lambda_{pk}^*|) \\
&\equiv T_1 + T_2 + \sum_{k=1}^K T_{3k}. \tag{30}
\end{aligned}$$

Let $\hat{\mathbf{u}}_p = \underset{\mathbf{u}_p}{\operatorname{argmin}}(\mathcal{V}(\mathbf{u}_p) - \mathcal{V}(0))$, then $\lambda_p^{\text{lla}} = \lambda_p^* + \hat{\mathbf{u}}_p/\sqrt{N}$. When $N \rightarrow \infty$, $S_{yy} \rightarrow \Omega^* = \Lambda^* \Lambda^{*\top} + \Sigma^*$ element-wise in probability. If we represent $\Omega^{(0)} = \Lambda^{(0)} \Lambda^{(0)\top} + \Sigma$, $\Gamma^{(0)} = \Omega^{(0)-1} \Lambda^{(0)}$, and $\Psi^{(0)} = \mathbf{I}_K - \Lambda^{(0)\top} \Omega^{(0)-1} \Lambda^{(0)}$, then $\Psi^{(0)}$

$$\begin{aligned}
&= \Delta^{(0)} + \Lambda^{(0)\top} S_{yy} \Lambda^{(0)} \\
&= \mathbf{I}_K + \Lambda^{*\top} \left(\Omega^{(0)-1} S_{yy} \Omega^{(0)-1} - \Omega^{(0)-1} \right) \Lambda^* + \frac{\mathbf{u}^\top}{\sqrt{N}} \left(\Omega^{(0)-1} S_{yy} \Omega^{(0)-1} - \Omega^{(0)-1} \right) \Lambda^* + \\
&\quad \Lambda^{*\top} \left(\Omega^{(0)-1} S_{yy} \Omega^{(0)-1} - \Omega^{(0)-1} \right) \frac{\mathbf{u}}{\sqrt{N}} + \frac{\mathbf{u}^\top}{\sqrt{N}} \left(\Omega^{(0)-1} S_{yy} \Omega^{(0)-1} - \Omega^{(0)-1} \right) \frac{\mathbf{u}}{\sqrt{N}}. \tag{31}
\end{aligned}$$

Furthermore, $\Omega^{(0)-1} = \Omega^{*-1} + o_{\mathbb{P}}(1)$ by continuous mapping theorem and $S_{yy} = \Omega^* + o_{\mathbb{P}}(1)$; therefore,

$$\Psi^{(0)} = \mathbf{I}_K + o_{\mathbb{P}}(1) \left(\Lambda^{*\top} \Lambda^* + \frac{\mathbf{u}^\top}{\sqrt{N}} \Lambda^* + \Lambda^{*\top} \frac{\mathbf{u}}{\sqrt{N}} + \frac{\mathbf{u}^\top \mathbf{u}}{N} \right), \tag{32}$$

which in turn implies that $\Psi^{(0)} = \mathbf{I}_K + o_{\mathbb{P}}(1)$ because Λ^* is fixed and $\mathbf{u}_{pk} = o_{\mathbb{P}}(1)$ for $p = 1, \dots, P$ and $k = 1, \dots, K$. Then arguments similar to Theorem 2 of Zou (2006) and Theorem 5 of Zou and Li (2008) imply that

$$T_1 \equiv \frac{1}{2} \mathbf{u}_p^\top \frac{\Psi^{(0)}}{\sigma_{pp}^{2(0)}} \mathbf{u}_p \rightarrow \frac{1}{2} \frac{\mathbf{u}_p^\top \mathbf{u}_p}{\sigma_{pp}^{2(0)}}; \quad T_2 \equiv \frac{\sqrt{N} \mathbf{u}_p^\top (\hat{\lambda}_p^{(0)} - \Psi^{(0)} \lambda_p^*)}{\sigma_{pp}^{2(0)}} \rightarrow \frac{\mathbf{u}_p^\top \mathbf{l}_p}{\sigma_{pp}^{2(0)}} \tag{33}$$

respectively in probability and distribution, where $\mathbf{l}_p \sim \mathcal{N}(0, \mathbf{I}_K)$.

If $\lambda_{pk}^* \neq 0$ and $N \rightarrow \infty$, then $\lambda_{pk}^{(0)} \rightarrow \lambda_{pk}^*$ in probability due to \sqrt{N} -consistency of $\Lambda^{(0)}$, $\frac{\alpha_k + 1}{\sqrt{N}} \rightarrow 0$ due to Assumption (A2); $(\eta_k + |\lambda_{pk}^{(0)}|) \rightarrow \lambda_{pk}^*$ in probability due to Assumption (A2), \sqrt{N} -consistency of $\Lambda^{(0)}$, and continuous mapping theorem; and $\sqrt{N}(|\lambda_{pk}^* + \frac{\mathbf{u}_{pk}}{\sqrt{N}}| - |\lambda_{pk}^*|) = \operatorname{sign}(\lambda_{pk}^*) \mathbf{u}_{pk}$. By Slutsky's and

continuous mapping theorems,

$$T_{3k} \equiv \frac{\alpha_k + 1}{\sqrt{N}} \frac{1}{(\eta_k + |\lambda_{pk}^{(0)}|)} \sqrt{N} (|\lambda_{pk}^* + \frac{u_{pk}}{\sqrt{N}}| - |\lambda_{pk}^*|) \rightarrow 0 \quad (34)$$

in probability. If $\lambda_{pk}^* = 0$ and $N \rightarrow \infty$, then $\sqrt{N}\lambda_{pk}^{(0)} = \mathcal{O}_{\mathbb{P}}(1)$ due to \sqrt{N} -consistency of $\Lambda^{(0)}$, $\alpha_k \rightarrow \infty$ due to Assumption (A2); $\sqrt{N}(\eta_k + \lambda_{pk}^{(0)}) = \mathcal{O}_{\mathbb{P}}(1)$ due to Assumption (A2), \sqrt{N} -consistency of $\Lambda^{(0)}$, and continuous mapping theorem; and $\sqrt{N}(|\lambda_{pk}^* + \frac{u_{pk}}{\sqrt{N}}| - |\lambda_{pk}^*|) = |u_{pk}|$. By Slutsky's and continuous mapping theorems,

$$T_{3k} \equiv \frac{\alpha_k + 1}{\sqrt{N}(\eta_k + |\lambda_{pk}^{(0)}|)} \sqrt{N} (|\lambda_{pk}^* + \frac{u_{pk}}{\sqrt{N}}| - |\lambda_{pk}^*|) \rightarrow \begin{cases} 0 & \text{if } u_{pk} = 0, \\ \infty & \text{otherwise} \end{cases} \quad (35)$$

in probability. Again, by Slutsky's theorem, $\mathcal{V}(u_p) - \mathcal{V}(0) \rightarrow \mathcal{V}^*(u_p)$ in distribution (30), where

$$\mathcal{V}^*(u_p) = \begin{cases} \frac{u_{\mathcal{A}_p}^T I_{|\mathcal{A}_{pN}}| u_{\mathcal{A}_{pN}}}{2\sigma_{pp}^{2(0)}} - \frac{u_{\mathcal{A}_{pN}}^T l_{p,\mathcal{A}_{pN}}}{\sigma_{pp}^{2(0)}} & \text{if } u_{pk} = 0 \text{ for all } k \notin \mathcal{A}_{pN}, \\ \infty & \text{otherwise.} \end{cases} \quad (36)$$

$\mathcal{V}(u_p) - \mathcal{V}(0)$ is convex, and the unique minimum of $\mathcal{V}^*(u_p)$ is $(l_{p,\mathcal{A}_p}, 0)$ for $p = 1, \dots, P$. Following the epi-convergence results of Geyer (1994) and Knight and Fu (2000), $u_{p,\mathcal{A}_{pN}} \rightarrow l_{p,\mathcal{A}_{pN}}$ in distribution and $u_{p,\mathcal{A}_{pN}^c} \rightarrow 0$ in distribution for $p = 1, \dots, P$. Because $l_{p,\mathcal{A}_{pN}}$ is distributed as $\mathcal{N}(0, I_{|\mathcal{A}_{pN}}|)$, the asymptotic normality of non-zero loadings is proved.

We now prove the consistency of λ_p^{lla} for $p = 1, \dots, P$. For $p = 1, \dots, P$ all $k \in \mathcal{A}_{pN}$, $\lambda_p^{\text{lla}} \rightarrow \lambda_p^*$ in probability; therefore, for all $k \in \mathcal{A}_{pN}$, $\mathbb{P}\{k \mid k \in \mathcal{A}_p^*\} \rightarrow 1$ in probability for $p = 1, \dots, P$. To prove the consistency of λ_p^{lla} , we finally show that for all $k' \notin \mathcal{A}_{pN}$, $\mathbb{P}\{k' \mid k' \in \mathcal{A}_p^*\} \rightarrow 0$ in probability for $p = 1, \dots, P$. For any p , assume that $k' \in \mathcal{A}_p^*$, then the necessary conditions of KKT optimality and the differential of $\mathcal{V}(u_p)$ (28) with respect to $\lambda_{pk'}$ imply that $NX_{k'}^{(0)T} (w_p^{(0)} - X^{(0)}\lambda_p^{\text{lla}}) = (\alpha_{k'} + 1)/(\eta_{k'} + |\lambda_{pk'}^{(0)}|)$. Using (35), we know that $(\alpha_{k'} + 1)/(\eta_{k'} + |\lambda_{pk'}^{(0)}|) 1/\sqrt{N} \rightarrow \infty$ in probability, and

$$\sqrt{N}X_{k'}^{(0)T} (w_p^{(0)} - X^{(0)}\lambda_p^{\text{lla}}) = X_{k'}^{(0)T} X^{(0)} \sqrt{N}(\lambda_p^* - \lambda_p^{\text{lla}}) + \sqrt{N}(X_{k'}^{(0)T} w_p^{(0)} - X_{k'}^{(0)T} X^{(0)}\lambda_p^*) \quad (37)$$

The first and second terms on the right hand side of (37) are asymptotically normal respectively due to the asymptotic normality of non-zero $\lambda_{pk}^{\text{lla}}$'s and (33). By Slutsky's theorem, the left hand side of (37) is also asymptotically normal; therefore, for $p = 1, \dots, P$,

$$\mathbb{P}\{k' \in \mathcal{A}_p^*\} \leq \mathbb{P}\left\{k' \mid X_{k'}^{(0)T} (w_p^{(0)} - X^{(0)}\lambda_p^{\text{lla}}) = \frac{\alpha_{k'} + 1}{\eta_{k'} + |\lambda_{pk'}^{(0)}|}\right\} \rightarrow 0 \quad (38)$$

in probability, which further implies that $k' \notin \mathcal{A}_{pN} \implies \mathbb{P}\{k' \in \mathcal{A}_p^*\} \rightarrow 0$ in probability. This proves the consistency of λ_{pk} s.

C.3 Proof of Theorem 4.2 (asymptotic normality and consistency of Σ^{lla})

To prove the consistency of Σ^{lla} , it is enough to prove that $\sigma_{\text{pp}}^{2^{(\text{t})}}$ (6) is consistent. For the \sqrt{N} -consistent sequence of estimators $\Lambda_{\text{p}}^{(0)}$, Assumption (A2) and continuous mapping theorem imply that if $N \rightarrow \infty$, then $\hat{\Lambda}^{(0)} = (\Omega^* + o_{\mathbb{P}}(1))(\Omega^{*-1}\Lambda^* + o_{\mathbb{P}}(1))$, $\lambda_{\text{p}}^{(0)} = \lambda_{\text{p}}^* + o_{\mathbb{P}}(1)$, and

$$\begin{aligned}\sigma_{\text{pp}}^{2^{\text{lla}}} &= \lambda_{\text{p}}^{*\top} \lambda_{\text{p}}^* + o_{\mathbb{P}}(1) - 2\lambda_{\text{p}}^{*\top} \lambda_{\text{p}}^* + o_{\mathbb{P}}(1) + (\Omega^*)_{\text{pp}} + o_{\mathbb{P}}(1) \\ &= -\lambda_{\text{p}}^{*\top} \lambda_{\text{p}}^* + (\Omega^*)_{\text{pp}} + o_{\mathbb{P}}(1) = \sigma_{\text{pp}}^{2*} + o_{\mathbb{P}}(1),\end{aligned}\tag{39}$$

which proves the consistency of $\sigma_{\text{pp}}^{2^{\text{lla}}}$ and that of Σ^{lla} .

The asymptotic normality of Σ^{lla} follows from Theorem 5.21 of Van der Vaart (2000) due to the following two reasons. First, if $\sigma_{\text{pp}}^2 > 0$, then differential of (15) is convex in $1/\sigma_{\text{pp}}^2$ and its derivatives are continuous and bounded. This implies that the differential of (15) is locally Lipschitz in $1/\sigma_{\text{pp}}^2$ with a square integrable Lipschitz constant. Second, $\Sigma^{\text{lla}} \rightarrow \Sigma^*$ in probability; see also Theorem 12.1 of Anderson and Rubin (1956).

C.4 Bayesian model averaging for expandable factor analysis

We estimate $p(Y | \mathcal{M}_g)$ using integrated nested Laplace approximation. The main idea of the proof exploits the fact that if $\theta = (\text{vec}(\Lambda^T), \text{diag}(\Sigma))$, then $\mathcal{L}(\theta) \geq \mathcal{Q}_{\text{LLA}}(\theta | \theta^{\text{lla}})$ and $\mathcal{L}(\theta^{\text{lla}}) = \mathcal{Q}_{\text{LLA}}(\theta^{\text{lla}} | \theta^{\text{lla}})$; see (20) and (21). Laplace approximation at any grid point (ρ_g, δ_g) requires calculation of

$$\frac{\partial^2 \log p(Y, \Lambda | \theta^g)}{\partial \text{vec}(\Lambda^T) \partial \text{vec}(\Lambda^T)^T} = \frac{\partial^2 \log p(Y | \Lambda, \theta^g)}{\partial \text{vec}(\Lambda^T) \partial \text{vec}(\Lambda^T)^T} + \frac{\partial^2 \log p(\Lambda | \rho_g, \delta_g)}{\partial \text{vec}(\Lambda^T) \partial \text{vec}(\Lambda^T)^T},\tag{40}$$

for the non-zero entries of Λ . The left hand side of (40) is obtained using the right hand side. While the first term $\log p(Y | \Lambda, \Lambda^g, \Sigma^g)$ in the right hand side is differentiable, the term $\log p(\Lambda | \rho_g, \delta_g)$ is not. This problem is resolved by exploiting the Gaussian scale mixture representation of the multiscale generalized double Pareto prior $p(\Lambda | \rho_g, \delta_g)$ at θ^g so that

$$\log p(\lambda_{\text{p}} | \rho_g, \delta_g) \propto - \sum_{k \in \mathcal{A}_{\text{pN}}} \frac{\lambda_{\text{pk}}^2}{2} \frac{\alpha_k^g + 1}{|\lambda_{\text{pk}}^g|(\eta_k^g + |\lambda_{\text{pk}}^g|)},\tag{41}$$

where λ_{p} represents λ_{pk} with $k \in \mathcal{A}_{\text{pN}}$ to ease notation. Differentiating (41) twice with respect to λ_{p} yields

$$-\frac{\partial^2 \log p(\lambda_{\text{p}} | \rho_g, \delta_g)}{\partial \lambda_{\text{p}} \partial \lambda_{\text{p}}^T} = \text{diag} \left(\frac{\alpha_{k_1}^g + 1}{|\lambda_{\text{pk}_1}^g|(\eta_{k_1}^g + |\lambda_{\text{pk}_1}^g|)}, \dots, \frac{\alpha_{k_p}^g + 1}{|\lambda_{\text{pk}_p}^g|(\eta_{k_p}^g + |\lambda_{\text{pk}_p}^g|)} \right) \equiv D_{\text{p}}^g\tag{42}$$

for $k_1, \dots, k_p \in \mathcal{A}_{\text{pN}}$ and $p = 1, \dots, P$. The analytic form in (41) can be also obtained using local quadratic approximation, which has been used previously in penalized variable selection (Fan and Li, 2001; Hunter and Li, 2005). For $p = 1, \dots, P$, the p th block of the information matrix for $\text{vec}(\Lambda^T)$ evaluated at

θ^g

$$H_p^g = -\frac{\partial^2 \log p(Y, \Lambda \mid \theta^g)}{\partial \lambda_p \partial \lambda_p^T} = \frac{N \Psi^g}{\sigma_{pp}^{2g}} + D_p^g; \quad (43)$$

therefore, the overall information matrix for $\text{vec}(\Lambda^T)$

$$H = -\frac{\partial^2 \log p(Y, \Lambda \mid \theta^g)}{\partial \text{vec}(\Lambda^T) \partial \text{vec}(\Lambda^T)^T} = \text{bdiag}(H_1, \dots, H_p, \dots, H_p), \quad (44)$$

where bdiag is the block diagonal operator. Using (44), the analytic form of $\log p(Y \mid \mathcal{M}_g)$ follows from the standard definition of Laplace approximation

$$\log \pi_g \propto p(Y \mid \mathcal{M}_g) = \log p(Y \mid \Lambda^g, \Sigma^g) + \log p(\Lambda^g) + \frac{\log 2\pi}{2} \sum_{p=1}^P |\mathcal{A}_{pN}^g| - \frac{1}{2} \sum_{p=1}^P \log \left| \frac{N \Psi^g}{\sigma_{pp}^{2g}} + D_p^g \right|. \quad (45)$$

C.5 Proof of Theorem 4.3 (model selection consistency of expandable factor analysis)

Assume that $p(\mathcal{M}_g) = p_g$, where $0 < p_g < 1$ and $\sum_{g=1}^G p_g = 1$, and that $\pi_g = \mathbb{P}(\mathcal{M}_g \mid Y)$ represents the posterior probability that \mathcal{M}_g is true, then

$$\pi_g = \frac{p(Y \mid \mathcal{M}_g)p(\mathcal{M}_g)}{\sum_{g=1}^G p(Y \mid \mathcal{M}_g)p(\mathcal{M}_g)} \implies \pi_g = \frac{p(Y \mid \mathcal{M}_g)p_g}{\sum_{g=1}^G p(Y \mid \mathcal{M}_g)p_g}. \quad (46)$$

If index g^* corresponds to the true model, then

$$\pi_{g^*} = \mathbb{P}(\mathcal{M}_{g^*} \mid Y) = \left(1 + \sum_{g=1, g \neq g^*}^G \frac{p(Y \mid \mathcal{M}_g)p_g}{p(Y \mid \mathcal{M}_{g^*})p_{g^*}} \right)^{-1}. \quad (47)$$

To prove the model selection consistency of expandable factor analysis, we need to show that when $N \rightarrow \infty$,

$$\pi_{g^*} \rightarrow 1 \iff \frac{p(Y \mid \mathcal{M}_g)}{p(Y \mid \mathcal{M}_{g^*})} \rightarrow 0 \quad (48)$$

in probability for $g \neq g^*$; therefore, it is enough to show that $\log p(Y \mid \mathcal{M}_{g^*}) - \log p(Y \mid \mathcal{M}_g) \rightarrow \infty$ in probability as $N \rightarrow \infty$. The overfitted expandable factor analysis model that corresponds to grid index g

has $\mathcal{A}_{p_N}^{g*} \subset \mathcal{A}_{p_N}^g$ and its number of factors $K > K^*$. Using (45),

$$\begin{aligned} \log p(Y | \mathcal{M}_{g^*}) - \log p(Y | \mathcal{M}_g) &= \log \frac{p(Y | \Lambda^{g^*}, \Sigma^{g^*})}{p(Y | \Lambda^g, \Sigma^g)} + \log \frac{p(\Lambda^{g^*})}{p(\Lambda^g)} + \frac{1}{2} \sum_{p=1}^P \log \frac{\left| \frac{N\Psi^g}{\sigma_{pp}^{2g}} + D_p^g \right|}{\left| \frac{N\Psi^{g^*}}{\sigma_{pp}^{2g^*}} + D_p^{g^*} \right|} \\ &\quad - \frac{\log 2\pi}{2} \sum_{p=1}^P \left(|\mathcal{A}_{p_N}^{g*}| - |\mathcal{A}_{p_N}^g| \right) \\ &\equiv T_1 + T_2 + T_3 + T_4. \end{aligned} \quad (49)$$

The log likelihood of expandable factor analysis depends on the data only through S_{yy} , so it is bounded for any Λ^g, Σ^g $g = 1, \dots, G$. Also, Ω^g is estimated as a positive definite matrix, so $L_1 \leq |T_1| \leq L_2$, where L_1 and L_2 are fixed constants. We now find the stochastic order of T_4 using (30) and (33). Since Λ^g and Λ^{g^*} are estimated using \sqrt{N} -consistent estimators of Λ^* , so

$$\sum_{p=1}^P \left(|\mathcal{A}_{p_N}^{g*}| - |\mathcal{A}_{p_N}^g| \right) = P(K^* - K) \mathcal{O}_{\mathbb{P}} \left(\frac{1}{\sqrt{N}} \right) = o_{\mathbb{P}}(1). \quad (50)$$

We now obtain a lower bound for T_2 . The form of the multiscale generalized double Pareto prior implies that

$$\begin{aligned} T_2 &= - \sum_{p=1}^P \sum_{k \in \mathcal{A}_{p_N}^g \setminus \mathcal{A}_{p_N}^{g*}} \log \frac{\alpha_{k_N}^g}{2\eta_{k_N}^g} + \sum_{p=1}^P \sum_{k \in \mathcal{A}_{p_N}^{g*}} \log \frac{\alpha_{k_N}^{g*} \eta_{k_N}^g}{\alpha_{k_N}^g \eta_{k_N}^{g*}} + \\ &\quad \sum_{p=1}^P \left\{ \sum_{k \in \mathcal{A}_{p_N}^g} (\alpha_{k_N}^g + 1) \log \left(1 + \frac{|\lambda_{pk}^g|}{\eta_{k_N}^g} \right) - \sum_{k \in \mathcal{A}_{p_N}^{g*}} (\alpha_{k_N}^{g*} + 1) \log \left(1 + \frac{|\lambda_{pk}^{g*}|}{\eta_{k_N}^{g*}} \right) \right\} \\ &= - \sum_{p=1}^P \sum_{k \in \mathcal{A}_{p_N}^g \setminus \mathcal{A}_{p_N}^{g*}} \log \frac{\alpha_{k_N}^g}{2\eta_{k_N}^g} + \sum_{p=1}^P \sum_{k \in \mathcal{A}_{p_N}^{g*}} \log \frac{\alpha_{k_N}^{g*} \eta_{k_N}^g}{\alpha_{k_N}^g \eta_{k_N}^{g*}} + \\ &\quad \sum_{p=1}^P \sum_{k \in \mathcal{A}_{p_N}^g \setminus \mathcal{A}_{p_N}^{g*}} (\alpha_{k_N}^g + 1) \log \left(1 + \frac{|\lambda_{pk}^g|}{\eta_{k_N}^g} \right) + \\ &\quad \sum_{p=1}^P \sum_{k \in \mathcal{A}_{p_N}^{g*}} \left\{ (\alpha_{k_N}^g + 1) \log \left(1 + \frac{|\lambda_{pk}^g|}{\eta_{k_N}^g} \right) - (\alpha_{k_N}^{g*} + 1) \log \left(1 + \frac{|\lambda_{pk}^{g*}|}{\eta_{k_N}^{g*}} \right) \right\}. \end{aligned} \quad (51)$$

Because \mathcal{M}_g is overfitted in the number of factors, $\alpha_{k_N}^{g*} > \alpha_{k_N}^g$ for all k , $\eta^{g*} \leq \eta_{k_N}^g$ for all k , and

$\mathcal{A}_{p_N}^{g*} \subset \mathcal{A}_{p_N}^g$ for all p . Using this, (51) reduces to

$$\begin{aligned}
T_2 &\geq - \sum_{p=1}^P \sum_{k \in \mathcal{A}_{p_N}^g \setminus \mathcal{A}_{p_N}^{g*}} \log \frac{\alpha_{k_N}^g}{2\eta_{k_N}^g} + \sum_{p=1}^P \sum_{k \in \mathcal{A}_{p_N}^g \setminus \mathcal{A}_{p_N}^{g*}} (\alpha_{k_N}^g + 1) \log \left(1 + \frac{|\lambda_{pk}^g|}{\eta_{k_N}^g} \right) \\
&\quad + \sum_{p=1}^P \sum_{k \in \mathcal{A}_{p_N}^{g*}} (\alpha_{k_N}^g + 1) \left\{ \log \left(1 + \frac{|\lambda_{pk}^g|}{\eta_{k_N}^g} \right) - \log \left(1 + \frac{|\lambda_{pk}^{g*}|}{\eta_{k_N}^{g*}} \right) \right\} \\
&= - \sum_{p=1}^P \sum_{k \in \mathcal{A}_{p_N}^g \setminus \mathcal{A}_{p_N}^{g*}} \left\{ \log \frac{\sqrt{N}\alpha_{k_N}^g}{2\sqrt{N}\eta_{k_N}^g} + (\alpha_{k_N}^g + 1) \log \left(1 + \frac{\sqrt{N}|\lambda_{pk}^g|}{\sqrt{N}\eta_{k_N}^g} \right) \right\} \\
&\quad + \sum_{p=1}^P \sum_{k \in \mathcal{A}_{p_N}^{g*}} (\alpha_{k_N}^g + 1) \log \left(\frac{\sqrt{N}\eta_{k_N}^g + \sqrt{N}|\lambda_{pk}^g|}{\sqrt{N}\eta_{k_N}^{g*} + \sqrt{N}|\lambda_{pk}^{g*}|} \frac{\eta_{k_N}^{g*}}{\eta_{k_N}^g} \right) \\
&\geq - \sum_{p=1}^P \sum_{k \in \mathcal{A}_{p_N}^g \setminus \mathcal{A}_{p_N}^{g*}} \left\{ \log \frac{\sqrt{N}\alpha_{k_N}^g}{2\sqrt{N}\eta_{k_N}^g} + (\alpha_{k_N}^g + 1) \log \left(1 + \frac{\sqrt{N}|\lambda_{pk}^g|}{\sqrt{N}\eta_{k_N}^g} \right) \right\} \\
&\quad + \sum_{p=1}^P \sum_{k \in \mathcal{A}_{p_N}^{g*}} (\alpha_{k_N}^g + 1) \log \left(\frac{\sqrt{N}\eta_{k_N}^{g*} + \sqrt{N}|\lambda_{pk}^{g*}|}{\sqrt{N}\eta_{k_N}^{g*} + \sqrt{N}|\lambda_{pk}^g|} \right). \tag{52}
\end{aligned}$$

Assumption (A2) of Theorem 4.2 implies that $\alpha_{k_N}^g/\sqrt{N} \rightarrow 0$ and $\sqrt{N}\eta_{k_N}^g \rightarrow c_k > 0$ for all k and g as $N \rightarrow \infty$; Theorem 4.2 implies that $\lambda_{pk}^g = \lambda_{pk}^* + o_{\mathbb{P}}(1)$, $\lambda_{pk}^{g*} = \lambda_{pk}^* + o_{\mathbb{P}}(1)$; and Lemma 2.1 proves that $\{\alpha_{k_N}^g/\eta_{k_N}^g\}_{k=1}^\infty$ is an increasing sequence for all g ; therefore, (52) reduces to

$$\begin{aligned}
T_2 &\geq - \left\{ \log \frac{\sqrt{N}\alpha_{k_N}^g}{2c_1 + o_{\mathbb{P}}(1)} + (\alpha_{k_N}^g + 1) \log \left(1 + \frac{o_{\mathbb{P}}(1)}{c_1} \right) \right\} \sum_{p=1}^P (|\mathcal{A}_{p_N}^g| - |\mathcal{A}_{p_N}^{g*}|) \\
&\quad + (\alpha_{k_N}^g + 1) \left(\sum_{p=1}^P |\mathcal{A}_{p_N}^{g*}| + o_{\mathbb{P}}(1) \right) \\
&= - \left\{ \log \sqrt{N} + \log \alpha_{k_N}^g + \alpha_{k_N}^g \log \left(1 + \frac{o_{\mathbb{P}}(1)}{c_1} \right) + o_{\mathbb{P}}(1) \right\} P(K - K^*) o_{\mathbb{P}} \left(\frac{1}{\sqrt{N}} \right) \\
&\quad + (\alpha_{k_N}^g + 1) \left(\sum_{p=1}^P |\mathcal{A}_{p_N}^{g*}| + o_{\mathbb{P}}(1) \right) \\
&= (\alpha_{k_N}^g + 1) \text{const.} + o_{\mathbb{P}}(1) \xrightarrow{N \rightarrow \infty} \infty \tag{53}
\end{aligned}$$

in probability as $N \rightarrow \infty$, where $\text{const.} > 1$.

Finally, we obtain a lower bound for T_3 . We use the assumption that k th eigenvalue of $\Psi^{g*} e_k^{g*} > 0$ for

all k and $\sigma_{pp}^{2*} > 0$ for all p . Using (49),

$$\begin{aligned}
2T_3 &= \sum_{p=1}^P \log \frac{\left| \frac{N\Psi_p^g}{\sigma_{pp}^{2g}} + D_p^g \right|}{\left| \frac{N\Psi_p^{g*}}{\sigma_{pp}^{2g*}} + D_p^{g*} \right|}} \\
&= \sum_{p=1}^P \left\{ \sum_{k \in \mathcal{A}_{pN}^g} \log \left(\frac{Ne_k^g}{\sigma_{pp}^{2g}} + \frac{\alpha_{kN}^g + 1}{|\lambda_{pk}^g|(\eta_{kN}^g + |\lambda_{pk}^g|)} \right) - \sum_{k \in \mathcal{A}_{pN}^{g*}} \log \left(\frac{Ne_k^{g*}}{\sigma_{pp}^{2g*}} + \frac{\alpha_{kN}^{g*} + 1}{|\lambda_{pk}^{g*}|(\eta_{kN}^{g*} + |\lambda_{pk}^{g*}|)} \right) \right\} \\
&= \sum_{p=1}^P \left(|\mathcal{A}_{pN}^g| - |\mathcal{A}_{pN}^{g*}| \right) \log N + \sum_{p=1}^P \sum_{k \in \mathcal{A}_{pN}^g \setminus \mathcal{A}_{pN}^{g*}} \log \left(\frac{e_k^g}{\sigma_{pp}^{2g}} + \frac{(\alpha_{kN}^g + 1)}{|\lambda_{pk}^g| \sqrt{N} (\sqrt{N} \eta_{kN}^g + \sqrt{N} |\lambda_{pk}^g|)} \right) + \\
&\quad \sum_{p=1}^P \sum_{k \in \mathcal{A}_{pN}^{g*}} \left\{ \log \left(\frac{e_k^g}{\sigma_{pp}^{2g}} + \frac{(\alpha_{kN}^g + 1)/\sqrt{N}}{|\lambda_{pk}^g| (\sqrt{N} \eta_{kN}^g + \sqrt{N} |\lambda_{pk}^g|)} \right) - \log \left(\frac{e_k^{g*}}{\sigma_{pp}^{2g*}} + \frac{(\alpha_{kN}^{g*} + 1)/\sqrt{N}}{|\lambda_{pk}^{g*}| (\sqrt{N} \eta_{kN}^{g*} + \sqrt{N} |\lambda_{pk}^{g*}|)} \right) \right\}.
\end{aligned} \tag{54}$$

Theorem 4.2 implies that there exists $d_{pk} > 0$ such that $|\lambda_{pk}| = d_{pk}/\sqrt{N}$ for $k \in \mathcal{A}_{pN}^g \setminus \mathcal{A}_{pN}^{g*}$ and $|\mathcal{A}_{pN}^g| - |\mathcal{A}_{pN}^{g*}| \geq 1$ for all p . Using the same arguments involving Assumption (A2) of Theorem 4.2 and those used in deriving the lower bound for T_2 ,

$$\begin{aligned}
2T_3 &\geq \sum_{p=1}^P \left(|\mathcal{A}_{pN}^g| - |\mathcal{A}_{pN}^{g*}| \right) \left\{ \log N + \sum_{k \in \mathcal{A}_{pN}^g \setminus \mathcal{A}_{pN}^{g*}} \log \left(\frac{e_k^g}{\sigma_{pp}^{2g}} + \frac{\sqrt{N}(\alpha_{1N}^g + 1)}{d_{pk}(c_k + d_{pk}/\sqrt{N})} \right) \right\} + \\
&\quad \sum_{p=1}^P \sum_{k \in \mathcal{A}_{pN}^{g*}} \left\{ \log \left(\frac{e_k^g}{\sigma_{pp}^{2g} + o_{\mathbb{P}}(1)} + \frac{o(1)}{|\lambda_{pk}^*|(c_k + \sqrt{N}|\lambda_{pk}^*|) + o_{\mathbb{P}}(1)} \right) \right. \\
&\quad \left. - \log \left(\frac{e_k^{g*}}{\sigma_{pp}^{2g*} + o_{\mathbb{P}}(1)} + \frac{o(1)}{|\lambda_{pk}^{g*}|(c_k + \sqrt{N}|\lambda_{pk}^{g*}|) + o_{\mathbb{P}}(1)} \right) \right\} \\
&= P \left\{ \log N + \sum_{k \in \mathcal{A}_{pN}^g \setminus \mathcal{A}_{pN}^{g*}} \log \left(\frac{e_k^g}{\sigma_{pp}^{2g}} + \frac{\sqrt{N}(\alpha_{1N}^g + 1)}{d_{pk}(c_k + d_{pk}/\sqrt{N})} \right) \right\} \\
&\quad \sum_{p=1}^P \sum_{k \in \mathcal{A}_{pN}^{g*}} \log \left(1 + \frac{o_{\mathbb{P}}(1)}{e_k^{g*} |\lambda_{pk}^*|(c_k + \sqrt{N}|\lambda_{pk}^*|)} \right) + o_{\mathbb{P}}(1) \\
&= \mathcal{O}(\log N) + \mathcal{O}_{\mathbb{P}} \left(\log \alpha_{kN}^g \right) + o_{\mathbb{P}}(1) \longrightarrow \infty
\end{aligned} \tag{55}$$

in probability as $N \longrightarrow \infty$.

Combining (50), (53), and (55) implies that

$$\log p(Y | \mathcal{M}_{g*}) - \log p(Y | \mathcal{M}_g) \longrightarrow \infty \tag{56}$$

in probability as $N \longrightarrow \infty$.

Assumption (A2) of Theorem 4.2 implies that $\log p(Y | \mathcal{M}_{g^*}) - \log p(Y | \mathcal{M}_g) \rightarrow \infty$ in probability as $N \rightarrow \infty$ for all $g \neq g^*$. This proves the model selection consistency of expandable factor analysis, i.e., $p(Y | \mathcal{M}_g)/p(Y | \mathcal{M}_{g^*}) \rightarrow 0$ in probability for all $g \neq g^*$. A simple consequence of Theorems 4.2 and 4.3 is that the model averaged estimate of the overfitted expandable factor analysis model $\bar{\Lambda}^{\text{lla}} = \sum_g^G \pi_g \Lambda^g \rightarrow \Lambda^*$ in probability as $N \rightarrow \infty$.

References

- Ahn, S. C. and A. R. Horenstein (2013). Eigenvalue ratio test for the number of factors. *Econometrica* 81(3), 1203–1227.
- Anderson, T. W. and H. Rubin (1956). Statistical inference in factor analysis. In *Proceedings of the Third Berkeley Symposium on Mathematical Statistics and Probability*, Volume 5, pp. 111–150.
- Armagan, A., D. B. Dunson, and J. Lee (2013). Generalized double Pareto shrinkage. *Statistica Sinica* 23(1), 119–143.
- Bai, J. and K. Li (2012). Statistical analysis of factor models of high dimension. *The Annals of Statistics* 40(1), 436–465.
- Bhattacharya, A. and D. B. Dunson (2011). Sparse Bayesian infinite factor models. *Biometrika* 98(2), 291–306.
- Brown, C. D., L. M. Mangravite, and B. E. Engelhardt (2013). Integrative modeling of eqtls and cis-regulatory elements suggests mechanisms underlying cell type specificity of eqtls. *PLoS Genetics* 9(8), e1003649.
- Candes, E. J., M. B. Wakin, and S. P. Boyd (2008). Enhancing sparsity by reweighted ℓ_1 minimization. *Journal of Fourier analysis and applications* 14(5-6), 877–905.
- Carvalho, C., J. Chang, J. Lucas, J. Nevins, Q. Wang, and M. West (2008). High-dimensional sparse factor modeling: applications in gene expression genomics. *Journal of the American Statistical Association* 103(484), 1438–1456.
- Chen, J. and Z. Chen (2008). Extended Bayesian information criteria for model selection with large model spaces. *Biometrika* 95(3), 759–771.
- Dempster, A. P., N. M. Laird, and D. B. Rubin (1977). Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society B* 39(1), 1–38.
- Fan, J. and R. Li (2001). Variable selection via nonconcave penalized likelihood and its oracle properties. *Journal of the American Statistical Association* 96(456), 1348–1360.
- Geyer, C. J. (1994). On the asymptotics of constrained M-estimation. *The Annals of Statistics*, 1993–2010.

- Golub, G. H. and C. F. Van Loan (2012). *Matrix Computations*, Volume 3. JHU Press.
- Hirose, K. and M. Yamamoto (2014). Estimation of an oblique structure via penalized likelihood factor analysis. *Computational Statistics & Data Analysis* 79, 120–132.
- Hunter, D. R. and R. Li (2005). Variable selection using MM algorithms. *The Annals of Statistics* 33(4), 1617–1642.
- Jolliffe, I. T., N. T. Trendafilov, and M. Uddin (2003). A modified principal component technique based on the Lasso. *Journal of Computational and Graphical Statistics* 12(3), 531–547.
- Knight, K. and W. Fu (2000). Asymptotics for Lasso-type estimators. *The Annals of Statistics*, 1356–1378.
- Knowles, D. and Z. Ghahramani (2011). Nonparametric Bayesian sparse factor models with application to gene expression modeling. *The Annals of Applied Statistics* 5(2B), 1534–1552.
- Lam, C. and Q. Yao (2012). Factor modeling for high-dimensional time series: inference for the number of factors. *The Annals of Statistics* 40(2), 694–726.
- Mazumder, R., J. H. Friedman, and T. Hastie (2011). Sparsenet: Coordinate descent with nonconvex penalties. *Journal of the American Statistical Association* 106(495), 1125–1138.
- Pati, D., A. Bhattacharya, N. S. Pillai, and D. B. Dunson (2014). Posterior contraction in sparse Bayesian factor models for massive covariance matrices. *The Annals of Statistics* 42(3), 1102–1130.
- Rubin, D. B. and D. T. Thayer (1982). EM algorithms for ML factor analysis. *Psychometrika* 47(1), 69–76.
- Rue, H., S. Martino, and N. Chopin (2009). Approximate Bayesian inference for latent Gaussian models by using integrated nested Laplace approximations. *Journal of the Royal Statistical Society B* 71(2), 319–392.
- Sethuraman, J. (1994). A constructive definition of Dirichlet measures. *Statistica Sinica* 4, 639–650.
- Shen, H. and J. Z. Huang (2008). Sparse principal component analysis via regularized low rank matrix approximation. *Journal of Multivariate Analysis* 99(6), 1015–1034.
- Thorisson, G. A., A. V. Smith, L. Krishnan, and L. D. Stein (2005). The International HapMap project web site. *Genome Research* 15(11), 1592–1593.
- Van der Vaart, A. W. (2000). *Asymptotic Statistics*, Volume 3. Cambridge University Press.
- Witten, D. M., R. Tibshirani, and T. Hastie (2009). A penalized matrix decomposition, with applications to sparse principal components and canonical correlation analysis. *Biostatistics* 10(3), 515–534.
- Zhang, C.-H. (2010a). Nearly unbiased variable selection under minimax concave penalty. *The Annals of Statistics* 38(2), 894–942.

- Zhang, T. (2010b). Analysis of multi-stage convex relaxation for sparse regularization. *The Journal of Machine Learning Research* 11, 1081–1107.
- Zou, H. (2006). The adaptive Lasso and its oracle properties. *Journal of the American Statistical Association* 101(476), 1418–1429.
- Zou, H., T. Hastie, and R. Tibshirani (2006). Sparse principal component analysis. *Journal of Computational and Graphical Statistics* 15(2), 265–286.
- Zou, H. and R. Li (2008). One-step sparse estimates in nonconcave penalized likelihood models. *The Annals of Statistics* 36(4), 1509–1533.